Protein-ligand binding sites
Identification, characterization and interrelations

Peter Schmidtke

Departamento de Fisicoquímica
Facultad de Farmàcia
Universitat de Barcelona

Protein-ligand binding sites
Identification, characterization and interrelations

Peter Schmidtke
Septiembre 2011

Programa de Doctorado de Biomedicina, Universitat de Barcelona
Director de tesis: Dr. Xavier Barril Alonso

UNIVERSITAT DE BARCELONA

Departamento de Fisicoquímica
Facultad de Farmàcia
Universitat de Barcelona
Programa de Doctorado de Biomedicina

Protein-ligand binding sites
Identification, characterization and interrelations

Esta tesis ha sido realizada por el licenciado Peter Schmidtke bajo la dirección del Dr. Xavier Barril Alonso, Profesor de Investigación ICREA en el Departament del Departament de Fisicoquímica de la Facultad de Farmàcia de la Universitat de Barcelona, para optar al titulo de doctor por la Universitat de Barcelona en el Programa de Doctorado en Biomedicina.

Director de tesis	Doctorando
Dr. Xavier Barril Alonso	Peter Schmidtke

Der Zustand der gesamten menschlichen Moral lässt sich in zwei Sätzen zusammenfassen: We ought to. But we don't.

Kurt Tucholsky

Acknowledgements

Aquesta és probablement la part més personal del document, allà on puc dir el que em plagui, o gairebé. Han estat tres anys de treball, en un entorn increïble. He estat adoptat molt ràpidament per gent molt amable del meu entorn. Provenint d'un país famós per veure cervesa, i havent passat molt de temps en un país famós per veure vi, ara he tingut la oportunitat de descobrir un país famós per veure cava, Catalunya, Espanya.

Malgrat que no parlava ni mica de català o castellà, us les vareu arreglar per fer-me sentir immediatament còmode, intencionalment o no. Tot i que encara només puc parlar castellà com un francès que s'ha empassat un grapat de frankfurts, estic content de poder entendre la major part dels vostres idiomes, i això és gràcies a vosaltres.
En los parágrafos siguientes me gustaría agradecer a la COPPO (Comisión Organizadora de Pica-Pica Oficiales). Gracias Flavio por tus iniciativas repetidas y buen humor. Gracias a la madre del grupo, Assumpta por ser tan abierta y por su hospitalidad. Muchas gracias a Ramon por discusiones sobre temas inhabituales, Ana por discusiones interesantes, Carles y Jordi por su buen humor y bueno...el futbol. Gracias al chico que me llamó Tutututu, Patricio, y a Oscar por su hospitalidad.
Gracias a la chica de Xavi...Juana...eh...Montse, y a mi novio también. Por fin una experimentalista en el labo. Thanks to Mousumi, for countless nice Indian dishes, chats and working hours. Gracias al Ronald Mc Jesús por interminables discusiones sobre puntos de acuerdo (pero no muy claro). Gracias al Sergi por sus ánimos para acabar este trabajo. Me gustaría dar las gracias al Dani por su curiosidad y su mentalidad abierta. Ha sido un placer trabajar con todos vosotros! Fue una experiencia única que nunca olvidaré.
Muchas gracias a Marta, Pep, Natalia y Marc por distraer-me de mi trabajo con interminables partidos de openarena ;)

Je voudrais profiter de ce petit paragraphe entre tant d'autres pour remercier ma deuxième famille, ma famille française à savoir Sylvette, Maurice et ma Bloupette. C'est grace à votre aide constante et importante que j'ai pu avoir un aperçu d'une toute autre culture, d'un nouveau pays et d'une nouvelle langue. Merci de m'avoir fait confiance dans mes choix académiques et professionels.
Mille fois merci aussi à Aurélie, Cham, Julien, Manu et Marine pour m'écouter de manière intéressée quand je vous ennuie avec des choses scientifiques et surtout merci d'avoir gardé ma chérie avec soins durant mes trois années de thèse!
Un grand merci à Guilhem avec lequel c'est toujours divertissant de philosopher sur les aléas de la science. Un merci particulier est adressé à Vincent pour sa franchise et le bon travail qu'on a pu faire ensemble durant les 4 dernières années. Concernant les années à venir....voyons ;)

Einen speziellen Dank an meine Familie, ohne die ich heute nicht an dieser

Stelle wäre. Mutti, Papi, Martin, Alex und beide Omis und Opas, danke für meine schöne Kindheit, meine Erziehung und den Freiraum den ihr mir gelassen habt um meinen Zielen nachzugehen. Vielen Dank für 28 Jahre Unterstützung! Lieben Dank an Frau Wobst, dafür, dass Sie mein Interesse an Themen wie Molekularbiologie und Chemie geweckt haben. Vielen lieben Dank auch an Professor Rübsamen für seine Unterstützung während meiner akademischen Laufbahn.

Un grand merci est adressé à l'équipe du Master de bio-informatique de l'université Paris Diderot. En particulier je voudrais remercier Patrick Fuchs, Delphine Flatters, Catherine Etchebest et Alexandre de Brevern pour m'avoir donné la possibilité d'évoluer dans leur excellente formation. Merci mille fois pour l'introduction à la modélisation moléculaire!

Me gustaría dar las gracias a Víctor Guallar y Jordi Mestres por el tiempo y interés invertido y por las críticas constructivas durante la evaluación de este trabajo. Un agradecimiento muy especial a Javi. Siempre estás disponible, con opiniones claras y constructivas. Si tan sólo el mundo científico tuviese más Javis, muchos aspectos estarían mejor ;) El último a quien tengo que agradecer es a Xavi, el jefe...aunque no es un jefe normal. Te debo muchísimo y espero que este trabajo te haya dado algo de razón por darme confianza al principio y durante la tesis. Trabajar contigo es muy estimulante, porque eres un científico optimista, abierto a ideas nuevas. He disfrutado mucho estos últimos tres años! Gracias a todos!

Diese Doktorarbeit ist jedem gewidment, der sie lesen möchte.

Contents

Chapter 1

Introduction

Observe, learn, predict; these are the three main steps for most scientific discovery processes, especially in biological sciences. We gradually learn and accumulate knowledge about a subject of interest. Once this knowledge is sufficiently abundant we are likely to derive a theory, or a model explaining what we observe. Ultimately, these theories and models are used to predict and classify new observations. Among other famous examples of such a typical scientific process, the theory of evolution by Charles Darwin.

Modern biomedical research follows very similar principles on a different scale. It is a very active field at the interface between basic and advanced applicative research linking biological observables to medical outcomes. Several new experimental techniques appeared during the last century helping in this complex endeavour. However, two events changed modern biomedical science and our everyday life profoundly in the last century, the introduction of mainstream computing and the development of a global network between computers, the internet. Computers have been used since the 1960s in research, but initial cost was prohibitive and calculation power limited. Today, average smartphones have a peak performance that is around 10 times higher than a Cray-1, a 5.5 tons weighting supercomputer in the 1970's. If Darwin had wanted to discuss and share his notes with several researchers in different places in the world he would probably have lost precious years for doing actual investigation. Today, the internet allows to send tremendous amounts of data nearly instantaneously to other parts of the world making research globally accessible. Ultimately, very much like Darwin's notebook at the time, modern data storage solutions allow to gather and organise together an unimaginable amount of biological observables. In order to analyse this data, whole new scientific disciplines have seen the day, like bio-informatics, systems biology and chemo-genomics. These disciplines try to analyse, learn and interpret this data to derive theories and models to predict biological outcomes. The work presented in this thesis is situated on the intersection of these three disciplines.

The protein data bank

A data collection of particular importance for this work is the Protein Data Bank (PDB) [Berman et al., 2000]. In this data bank a multitude of macromolecular structures, mostly of biological interest, is stored and organised. The PDB was born in the 1970s with the appearance of several crystal structures of proteins. Crystallography had indeed made tremendous advances for solving protein 3D structures [Perutz, 1970]. Today, over 60.000 protein crystal structures are available in the PDB, accessible via the internet.
Such structures are very complex and difficult to analyse and still there is a substantial amount of information than can not be understood. This problem is very similar to fields like genomics. Improvements in sequencing techniques make is possible to retrieve a genome sequence of a whole organism very rapidly. However, the information is so complex and so abundant that analysing this data alone is a field of investigation on its own.

An important part of my work intends to analyse protein structures (observe), derive theories (learn) and use them to predict several properties on a global scale (predict) throughout the PDB. A particular focus is set on applications in structure based target discovery and characterisation for drug design purposes.

Drug discovery

In industrialized countries we are used to have access to medical treatments against diseases like hypertension, cancer or AIDS. What is less commonly known is the process, time and amount of work behind the little pill taken every day by some people. Usually more than ten years separate an initial discovery of a target to a final drug on the market. During this time several steps are more or less recursively followed.
Starting with a disease of interest, first steps intend to isolate possible targets, proteins, that either cause a disease or could alter the outcome of a phenotype linked to the disease. Once such a target is identified, validated and assays are set up, the search for small molecules begins. Several techniques, both experimental and computational are employed to discover so-called hits, small molecular prototypes showing weak potency to the investigated targets. These molecules are then further optimised to so-called leads. In pre-clinical stages the ADMET properties and pharmacokinetic profiles of the selected compounds are investigated in in-vitro and in-vivo assays. Finally, during clinical trials compounds are tested in humans for efficacy and toxicity before entering the market.
The work presented in this thesis is intimately linked to various of these steps in drug discovery. Two contributions in this thesis specially focus on the characterisation of proteins as pharmaceutical targets. A third work in this thesis is shown to have possible applications in hit-to-lead optimisation and lead optimisation processes. The last two contributions are related to a manifold of fields during the drug discovery process. Possible applications range from areas like promiscuity evaluations of a chosen target in the beginning of the drug discovery process, to the re-use of already existing drugs for different purposes, drug-repurposing.

Scope of the work in a global project

Several parts of the work presented here can be perceived as independent projects tackling various issues related to computational drug discovery. However, this work is also part of a larger project carried out in our research group. This project is guided by the following observation: Most computational drug discovery efforts focus on the identification of ligands binding to known active or binding sites via equilibrium binding. However, it is known that various successful drugs on the market do not use this precise molecular mechanism of action [Swinney and Anthony, 2011]. Indeed, other mechanisms of action can be considered.
Lately, pharmaceutical industry became more involved in the investigation of protein-protein complexes. The discovery of drug-like molecules inhibiting the association of proteins shows the potential of pursuing other strategies in drug

development [Ding et al., 2005].
In our group we are particularly interested in alternative strategies to target proteins, notably allosteric inhibition, and targeting protein-protein interactions. However, instead of inhibiting an association of two proteins, another possible strategy can be to stabilise the interaction. As a matter of fact, a few known small molecules, such as brefeldin A [Renault et al., 2003] and rapamycin [Liang et al., 1999] act as protein-protein stabilisers. Despite this observation protein-protein stabilisation is currently subject of very moderate research.
To prove that this strategy is of therapeutic interest and feasible, my task consisted in systematically mining the Protein Data Bank for protein-protein complexes that could potentially be stabilised via the action of a drug-like molecule. To do this, a new method had to be developed and a purpose built data-base was designed. Subsequently, the identified targets will be further analysed by other members of our research group using a sophisticated target assessment method developed in-house. Once the potential targets further validated with this method, virtual screening will be performed, docking a library of compounds to the potential binding sites. Last, the identified compounds are purchased and tested experimentally for their ability to stabilise the interaction between both protein partners.

Accessibility of research

A very important aspect in modern research is communication. This communication is mainly facilitated by the internet and permits to share information, new findings and ideas in a very short time.
Areas of research related to a drug discovery are close to applicative outcomes. For this reason industrial and financial interests hinder free circulation of tools, ideas and research results in this area. This behaviour is likely to slow down research and progress [Edwards et al., 2009].
A central aspect of the work presented in this thesis is related to its accessibility. Most of the software and tools are published as open-source software. Data is made available and can be visualised via other open-source software.

Bibliography

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

Ke Ding, Yipin Lu, Zaneta Nikolovska-Coleska, Su Qiu, Yousong Ding, Wei Gao, Jeanne Stuckey, Krzysztof Krajewski, Peter P Roller, York Tomita, Damon A Parrish, Jeffrey R Deschamps, and Shaomeng Wang. Structure-based design of potent non-peptide MDM2 inhibitors. *Journal of the American Chemical Society*, 127(29):10130–1, July 2005.

Aled M Edwards, Chas Bountra, David J Kerr, and Timothy M Willson. Open access chemical and clinical probes to support drug discovery. *Nature chemical biology*, 5 (7):436–40, July 2009.

J Liang, J Choi, and J Clardy. Refined structure of the FKBP12-rapamycin-FRB ternary complex at 2.2 A resolution. *Acta crystallographica. Section D, Biological crystallography*, 55(Pt 4):736–44, April 1999.

M F Perutz. Stereochemistry of cooperative effects in haemoglobin. *Nature*, 228 (5273):726–39, November 1970.

Louis Renault, Bernard Guibert, and Jacqueline Cherfils. Structural snapshots of the mechanism and inhibition of a guanine nucleotide exchange factor. *Nature*, 426 (6966):525–30, December 2003.

David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews. Drug discovery*, 10(7):507–19, January 2011.

Chapter 2

Materials & Methods

2.1 Proteins

Accounting for around 50% of the dry mass in Escherichia coli [Stouthamer, 1973], proteins play a major role in nature. Often designated as the factories of the cell, proteins do not only play a functional, but also a structural role.
The study of the protein structure revealed several recurrent motives and patterns. These are often classified into four hierarchical levels further explained in the following pages.

2.1.1 Amino acid

Amino acids are small molecules constituting the building blocks of proteins. Composed of an amine group and a carboxylic group they differ in the side chain they carry on the C_a atom. In total 20 amino acids are encoded in the genome. Through transcription and translation the information carried by DNA is transformed into a sequence of amino acids. Amino acids can polymerize by forming a peptide bond between the carbonyle group of one amino acid and the amide group of another.

2.1.2 Primary Structure

This first level of protein structure hierarchy is simply the sequence of amino acids polymerized to a chain of amino acids linked by so-called peptide bonds. Shorter chains of amino acids are usually called peptides. The term protein is reserved for polypeptide chains above a given size usually with a precise function. This structure can be a well defined three dimensional structure [Lodish et al., 2004], but can also contain disordered regions.
The main chain of the amino acid sequence is composed of atoms constituting the peptide bond and the C terminal and C_a atom. This ensemble of atoms is termed backbone, as they constitute the main scaffold of the protein allowing for its relative structural stability. As previously mentioned, attached to the backbone C_a are side chains that differ from one amino acid to another. These side chains are usually more flexible than the backbone.

2.1.3 Secondary Structure

The primary sequence of amino-acids is able to arrange itself in three dimensional structural patterns. These patterns are structurally stabilized by intramolecular hydrogen bonds. To form secondary structural motives, these hydrogen bonds occur only between backbone atoms, i.e. the oxygen of the peptide bond carbonyl and the nitrogen of the amide. The three different types of common secondary structure motives are described in the following paragraphs.

The α helix

An example of an alpha helix is shown on figure 2.1. The helical secondary structure is a periodic arrangement of hydrogen bonds formed between the carbonyl oxygen of residue (red) n and the amide nitrogen (blue) of residue

n+4 on the amino acid sequence. Thus the backbone adopts a spiral fold. α helices are frequently found in proteins, and they are the main constituent of trans-membrane domains. In 3D-visualization, secondary structure motives are often simplified. Instead of representing all atoms of the molecule a simplified representation is used. Here the semi transparent *cartoon* representation of the alpha helix is overlaid. Side chain atoms are shown as transparent sticks. In the cartoon representation an alpha helix is often represented as a flattened ribbon.

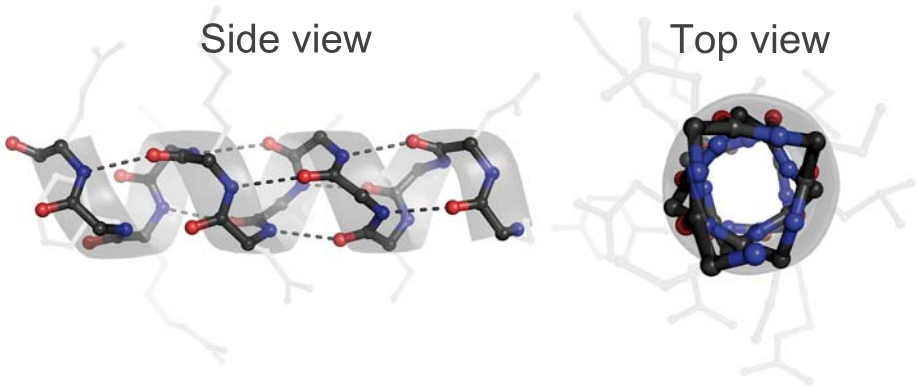


Figure 2.1: Example of an alpha helical secondary structure motif.

The β sheet

A β sheet is formed when 2 β strands (nearly linear extended backbone) are placed adjacent to each other allowing hydrogen bonding between the carbonyl oxygen and amide nitrogen of the backbone of the paired strand. β sheets are usually composed of strands of 5 residues or more. Figure 2.2 shows an example of a beta sheet. In the simplified cartoon representation a β strand is represented as a flat ribbon followed by an arrow indicating the direction of the amino-acid sequence. Beta sheets tend to be flat, but they can also show significant deformation to form for instance beta-barrels.

The turn

Turn motives allow very sharp redirections of the polypeptide chain. Usually built up by 3 to 4 residues, they are stabilized via one hydrogen bond again between the oxygen and nitrogen of two different residues (the end residues) of the turn. Proline and glycine residues are frequently found in turn motives. An example of a turn motif between 2 anti-parallel beta strands is shown on figure 2.3.

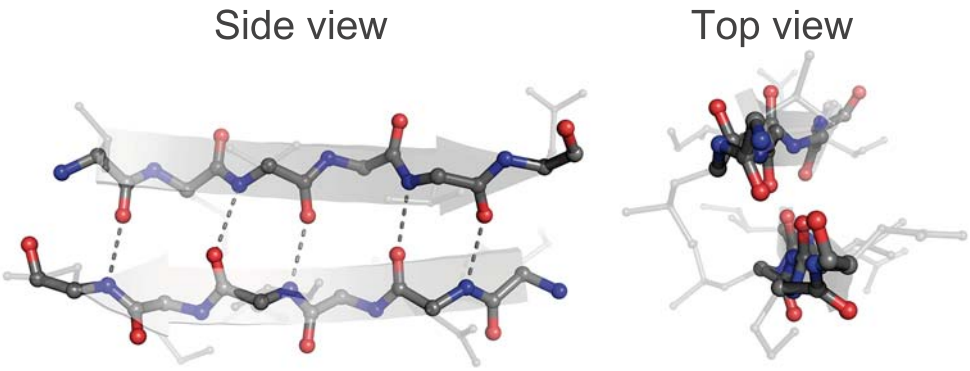


Figure 2.2: Example of a beta sheet.

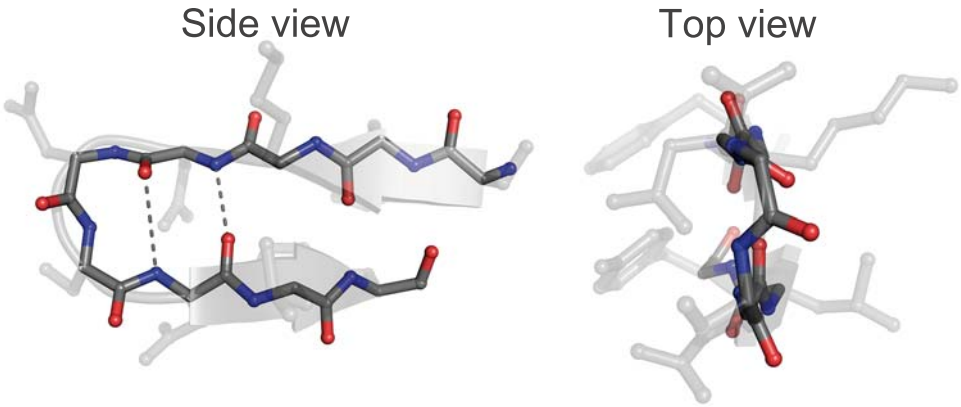


Figure 2.3: Example of a turn motif.

2.1.4 Tertiary Structure

The tertiary structure results of the way secondary structures arrange between each other to form the whole protein/domain. While secondary structural motives are mainly built up using hydrogen bonds as key interaction, tertiary structures are stabilised via a multitude of interactions. The main driving forces for globular protein folding are hydrophobic interactions, but ionic and hydrogen bonding interactions can also contribute to stabilize the structure. Furthermore, disulphide bridges can be built between cystein residues of a protein. Often, this level of hierarchy is the final 3D representation of a protein used in structural bio-informatics.

2.1.5 Quaternary Structure

The understanding, prediction and exploitation of the protein quaternary structure is another field of very active research. The quaternary structure corre-

sponds to the arrangement of two or more polypeptide chains to form a complex functional super-structure. In such a complex, the same peptide chain can be represented several times. In such a case the complex is designated as homo-multimer. If the peptides interacting in such a complex are different they are called hetero-multimers. A very well known example of a homo-dimer is HIV1-protease where two identical chains form the so-called biological unit of the protein. The biological unit describes the form of a protein that is able to fulfil a given function.

2.1.6 Intrinsically Disordered Proteins

The derivation of hierarchy in a protein structure is heavily based on observations made by experimental techniques like X-ray crystallography and NMR. However, these techniques imply that the protein is stable in a given conformational state. Lately, increasing interest is paid in understanding the role and structure of another category of proteins, loosely designated as intrinsically disordered proteins or IDP's. They are characterized by the absence of a stable secondary and tertiary structure which makes them difficult to observe experimentally. Nonetheless intrinsically disordered proteins play an important biological role [Tsvetkov et al., 2009]. While in this thesis the analysis is exclusively carried out on ordered proteins, it should be acknowledged that IDP's constitute an important separate class of proteins, which are building ground for a new exciting field of research.

2.1.7 Experimental methods for protein structure determination

The structure of a protein is crucial to its proper functioning. Knowledge of this three dimensional structure at an atomistic level can help us understanding structure function relationships, gain mechanistic insights into the role of the protein and last allow us to use this knowledge to alter or direct a proteins behaviour. Determining a protein structure is a long and expensive process. Currently, two main experimental methods are used to determine protein structure at an atomistic level, X-ray crystallography and NMR. Here basics of X-ray crystallography are shortly presented, because the work in this thesis relies substantially on structural information derived from crystal structures.

X-ray crystallography

Kendrew and Perutz were the pioneers of protein crystallography, publishing the first high resolution crystal structure of myoglobin in 1958. Since that date, 61683 protein crystals have been solved and deposited in the Protein Data Bank. Protein crystallography is a very long process, but this is not only due to methodologies used during structure determination itself. Several other steps have to be undertaken which are only shortly mentioned herein, but require substantial amount of time and work. First, the protein of interest has to be purified. Then, the protein has to be aligned in a rigid array to form a lattice of millions of repetitions of the same protein. This process is called

crystallization as this alignment results in a crystal, that can then further be analysed via techniques shown on figure 2.4. A beam of X-rays is sent through

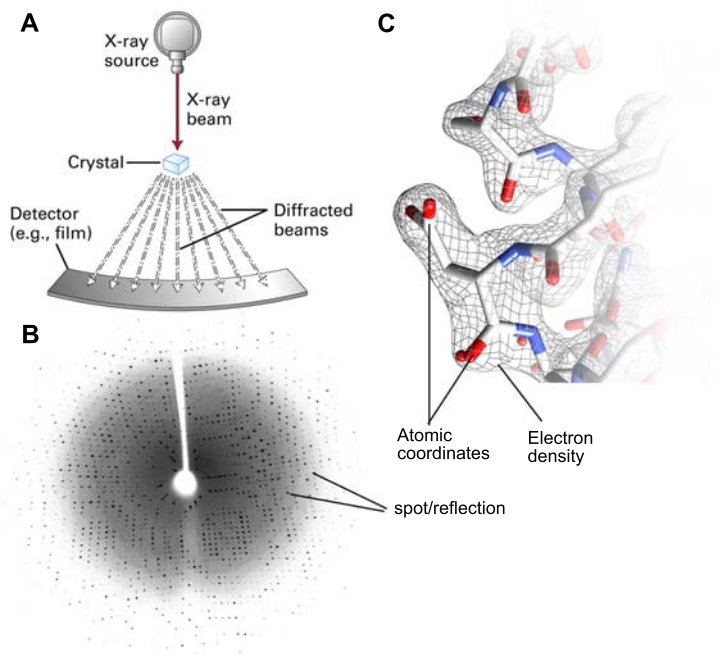


Figure 2.4: Schematic representation of the extraction of diffraction data via X-ray crystallography.

the crystal. The crystal scatters, or diffracts, the beam into a specific pattern of spots with varying intensities that can be captured via various means behind the crystal. Such a diffraction pattern is shown on figure 2.4 B. This pattern contains the relative information of atomic positions in space. To reconstruct the whole 3D structure, the crystal is rotated and several of these diffraction patterns are recorded.
Next, the series of diffraction patterns are converted to a 3 dimensional model of electron densities of the protein, as shown in figure 2.4 C. This process is mathematically very complex and involves different key steps. First each reflection is *indexed*, identifying which peak corresponds to which position in reciprocal space, accessible via Fourier transformation of the real space. Each rotation of the crystal during data collection produces one diffraction image. In a process called *merging* reflections from one image appearing in another image are identified. In a subsequent process called scaling, the relative intensities of the reflections between different diffraction images are optimised.
Practically, for calculating the electron density of a structure from a diffraction map, the indices of a reflection, the intensity of the spot and the phase angles of each reflection are required. While the first two can be derived directly from the diffraction plot, the phase angle is not known. Several techniques can be considered to determine the phase angle. These techniques won't be cov-

ered here, but more detail can be found in specialized introductory literature [Rhodes, 2006].
Once phases have been gathered, an initial atomic model of the molecule can be built into the electron density map. This model can then be used to calculate a theoretical diffraction pattern to verify how well the model fits the experimental data. Using various refinement cycles, the accordance between observed diffraction patterns and the final atomic model is optimised.
Two important parameters to assess the quality of a crystal and a final structural model are *(i)* the resolution expressed in $\mathring{A}$ and *(ii)* the R-value. The R-value expresses the accordance between the final model and experimental structure factor amplitudes derived from the diffraction pattern. The resolution of a structure is a global number expressing how uncertain the atom positions of the final model are on average, with high numbers corresponding to low resolution and vice-versa. For example, a resolution of $2.5\mathring{A}$ of a crystal structure yields uncertainties of around $0.4\mathring{A}$ on atom positions in the modelled structure.

Limitations While X-ray crystallography allowed the elucidation of an important number of macromolecular structures during the past decades, the technique has also limitations. Given the high impact the publication of the crystal structure of a protein has, these limitations are not always thoughtfully considered or even understood, resulting in publication of structures that are later found to be wrong (`http://retractionwatch.wordpress.com`).
A major limitation of X-ray crystallography is the fact that the protein has to be packed in a regular crystal. This process can alter the position of atoms in the protein and thus modify its shape [Eyal et al., 2005]. Furthermore, crystallization of the protein hinders observation of dynamics of the protein and provides a rather static image of the protein structure.
The construction of a molecular model into an electron density map is based on the fitting of atoms into such an electron density cloud. While proteins are built from a known sequence of amino-acids, which eases the structural prediction, small organic molecules are more difficult to identify only using the electron density, as are water molecules [Davis et al., 2003]. This can lead to misinterpretations of electron density maps and wrong ligand modelling into binding sites. Furthermore crystallization conditions can affect both protein and ligand conformation and position. This is particularly problematic for drug discovery, where the identification of the exact binding pose of a ligand in a binding site can give important insights into hotspots of the pocket and hints to further improve the ligand. Thus, before using a crystal structure for a particular project, uncertainties should be critically assessed.

2.1.8 Molecular surface

Three dimensional macromolecular structures are very complex in terms of shape and flexibility and various ways exist to describe it. While secondary structure representations help understanding the main architecture of a macromolecule an accurate estimation of the surface of a protein and an interaction

partner (macromolecule or small molecule) are crucial for understanding underlying concepts of molecular interaction.
Several computational means exist to approximate the molecular surface on static three dimensional structures. The three main approaches used in the field are resumed in the following subsections. Figure 2.5 shows schematically principles of each method.

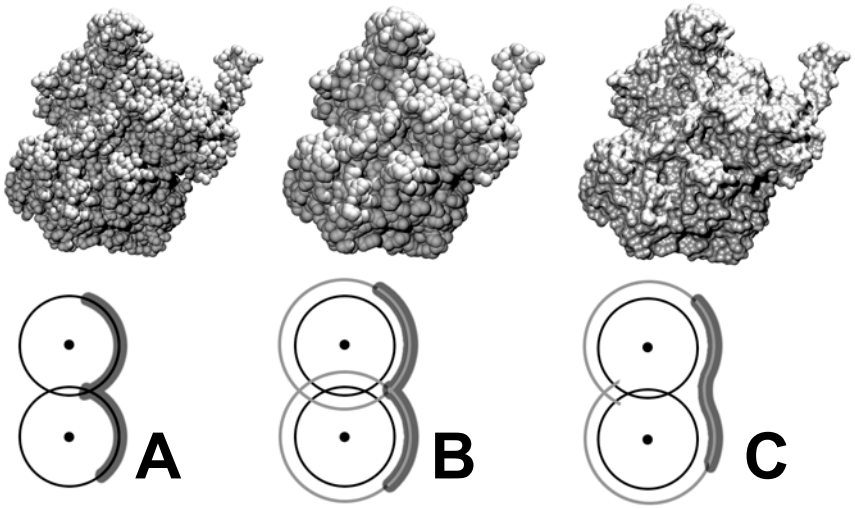


Figure 2.5: Different surface representations. **A:** Van der Waals surface, **B:** Lee-Richards surface, **C:** Connolly surface

Van der Waals surface

The van der Waals surface can be calculated by representing each atom by a sphere and the corresponding van der Waals radius of the atom. This representation is also called space-filling diagram and was introduced by [Lee and Richards, 1971]. It yields a multitude of protein internal interstitial volumes that are usually inaccessible for solvent molecules like water.

Lee-Richards surface or accessible surface

This surface representation is shown in figure 2.5 B. It is principally a van der Waals surface but using increased van der Waals radii by an increment corresponding to solvent radii [Lee and Richards, 1971]. For water molecules usually a radius of $1.4\mathring{A}$ is considered. Compared to the simple van der Waals surface, the accessible surface does not encompass internal interstitial surface patches. However, the surface still remains slightly over-evaluated, as the surface portion between two atoms remains inaccessible by another interacting atom.

Connolly surface or solvent accessible surface

The solvent accessible surface area is related to the accessible surface area and has been originally refined by Richards and Lee from their previous work on accessible surface area. Here, the surface is defined as being able to be contacted by a probe sphere of the size of a water molecule. Figure 2.5 C shows the principles behind the so-called reentrant surface, that is smoothed compared to accessible surface or van der Waals surface.

2.1.9 From structure to function

Figure 2.6 resumes schematically the multitude of functions a protein can have and that current research knows of. To achieve these functions, any given

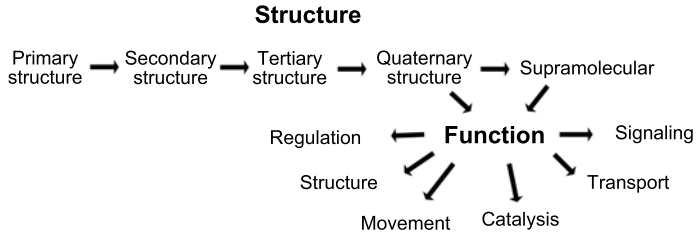


Figure 2.6: Overview of structure and function of proteins. Figure inspired by [Lodish et al., 2004])

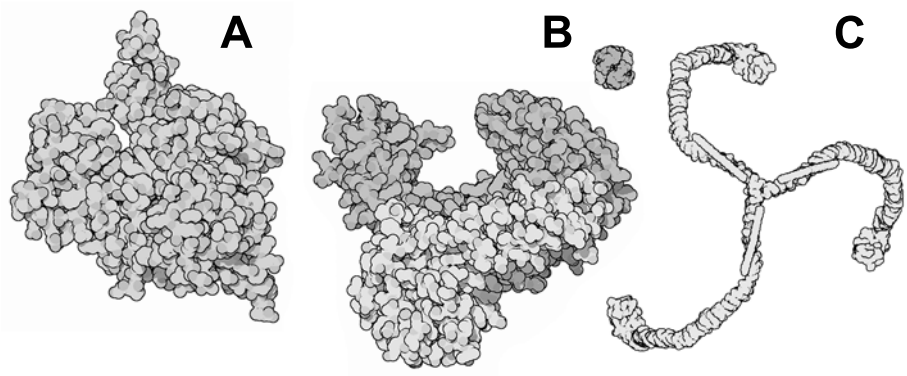


Figure 2.7: Examples of different protein shapes extracted from PDB-101. **A:** Actin, **B:** Reverse transcriptase, **C:** Clathrin

protein must carry out several complex tasks in a very specific manner. For instance, ion channel proteins must insert in the cell membrane, bind specifically to a signal molecule (e.g. a neurotransmiter), then allow certain ions to flow trough them. Additionally they must be susceptible to regulation by contact with other proteins or by posttranductional signals, etc. To achieve all of this, proteins must position the chemical functionalities of their residues

and prostetic groups in the manner necessary to attain the required physicochemical properties (shape, polarity, reactivity, flexibility, etc.) at the right 3D location, resulting in a wide variety of shapes as exemplified on figure 2.7.
Amongst the various functional units of proteins, small molecule binding sites are of particular interest because they are necessary for recognition of substrates by enzymes, hormones and other signalling molecules by receptors and allosterically regulated proteins, amongst other. The identification and characterisation of these sites is thus of interest to infer the function of a protein. Furthermore, they are privileged pharmacological sites of action. Section 2.3 explains in further detail methods to predict such small molecule binding sites and its implications.

2.2 Molecular simulations

Studying macromolecules like proteins, DNA or RNA on an atomic level using experimental techniques is very complex, time consuming and expensive. Furthermore, experimental structure determination techniques like X-ray crystallography can only give a limited insight into mechanistic details of a protein and even small molecules.
To ease the study of macromolecules at atomic resolution theoretical models can be used. The field of computational chemistry intends to develop new theoretical approaches and use them to *(i)* represent the molecule and *(ii)* simulate the molecule as accurately as possible.
In order to be able to study the dynamics of systems composed of thousands of atoms several approximations and techniques have been developed. These methods are less accurate than quantum mechanics but are commonly used to study macromolecular systems at an atomic level. In the following paragraphs principles of such an approximation (molecular mechanics) are very shortly outlined and applications to molecular dynamics (MD) and free energy calculations are shown.

2.2.1 Classical Mechanics & force fields

A fundamentally different approach to quantum chemistry is undertaken by classical molecular mechanics. Instead of describing electron positions around a nucleus, atoms are represented as a sphere of a given radius, charge and a mass. To calculate the potential energy of a system, charges, masses, radii, typical bond lengths, etcetera, have to be known for each atom and atom pairs. These parameters and corresponding equations using them constitute so-called force-fields. Several force-fields exist (Amber, Gromos, Charmm, MMFF, Dreiding ...) and mostly they are built on empirical knowledge or detailed calculations. Here I will focus on generalities on force-fields usually applied in simulating macromolecules. While all force-fields use different parameter sets and slightly different energy terms to calculate a systems' potential energy, globally all consider that the potential energy V of a system is additive and composed of the potential from bonded (or covalent) and non bonded (non covalent)

interactions of a system as represented by equation 2.1

$$V_{pot} = V_b + V_{nb} \tag{2.1}$$

where V_{pot} is the potential energy of a system and V_b the bonded interaction energy and V_{nb} the energy derived from non-bonded terms. Both, bonded and non-bonded terms are explained in more detail in the following paragraphs.

Bonded or covalent terms

This set of functions describes interactions between two atoms linked via a covalent bond. The covalent energy is composed of a sum of generally four terms, considered to be independent between each other.

$$V_b = V_{bond} + V_{angle} + V_{dihedral} + V_{improper} \tag{2.2}$$

Bond length: V_{bond} is a term estimating the energy related to the bond length between two atoms. As covalent bonds vibrate around a given bond length, harmonic potentials are used to approximate this behaviour.

$$V_{bond} = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 \tag{2.3}$$

Equation 2.3 is a form often used to calculate V_{bond}. Here the energy depends on the deviation of the observed bond length l_i from a reference (equilibrium or natural) bond length of the atom pair of $l_{i,0}$, where k_i is a force constant.

Bond angle: This term describes the angle observed between two adjacent bonds in a molecule. V_{angle} is very similar to the bond length. Here again a harmonic oscillator is used to estimate the energy corresponding to a deviation of the observed bond angle θ_i compared to the equilibrium bond angle $\theta_{i,0}$.

$$V_{angle} = \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 \tag{2.4}$$

Dihedral angles: This energy is calculated for bonds in the middle of 4 adjacent atoms. Generally the torsion angles formed by these 4 atoms can adopt a series of minima. Thus in the contrary to the previously used harmonic oscillators a more complex function is used to represent dihedrals.

$$V_{dihedral} = \sum_{torsions} \frac{V_n}{2}(1 + cos(n\omega - \gamma)) \tag{2.5}$$

where ω is the observed angle, γ the phase angle (indicating whether ω passes through a minimum). V_n is the height of the barrier and n the periodicity. The use of the cosine function results in a periodic energy function that can thus adopt various local minima.

Improper dihedral angles: Describing torsional angles is not enough to accurately describe a molecule. Often a last function estimating energy from improper dihedral angles is necessary. Typically, improper dihedrals are used to describe the planarity of a molecule or to prevent interconversion of stereocenters.

$$V_{improper} = \sum_{torsions} K_i(1 + cos(2\xi_n - 180)) \tag{2.6}$$

where ξ_n is the dihedral angle, K_i is a force constant.

Non-bonded or non-covalent terms

To describe long range and short range interactions between two atoms two energy terms are commonly used, the van der Waals energy and the electrostatic energy.

$$V_{nb} = V_{ele} + V_{vdw} \tag{2.7}$$

where V_{ele} is the electrostatic energy and V_{vdw} the van der Waals energy.

Electrostatic energy Contrary to quantum mechanics which explicitly accounts for electron distribution and nuclear charges, molecular mechanics approximate the electrostatic effect by means of partial point charges assigned to each atom centre. Due to this fact, Coulombs law can be used to estimate the electrostatic energy of two interacting molecules.

$$V_{ele} = \sum_{i=1} \sum_{j=1} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.8}$$

where q_i and q_j are point charges of atoms i and j, ε_0 is the dielectric constant of the medium and r_{ij} is the euclidean distance between both atoms. The most exact way to represent the solvent is to consider each solvent molecule explicitly. When molecular simulations with explicit solvent molecules are run, ε_0 is considered to be 1. Unfortunately, explicit solvent simulations are very costly, thus in so-called implicit solvent simulations the medium is approximated via the dielectric constant. Several approximations exist to estimate the effect of the solvent implicitly, like Generalized Born (GB), Linear Poisson-Boltzmann and non-linear Poisson Boltzmann. However, these are a crude approximation to represent the effect of a medium on an electrostatic interaction. In reality the dielectric constant is depending on the medium and an interaction occurring in bulk solvent would not require the use of the same dielectric constant as if it was happening on the surface of the protein. To account for the effect of the solvent on the electrostatic energy term, several other corrections can be used, like distance dependent dielectric constants for example.

Van der Waals energy Very short range interaction energies are highly influenced by van der Waals terms. These terms describe the behaviour of two atoms approaching each other without forming a covalent bond. When both atoms get too close to each other (inter-penetrate each other) a steric clash is modelled by large repulsive energies. Also an attractive term is included to

simulate London dispersion forces. The function found to reproduce such a behaviour is the Lennard Jones potential. While other descriptions of the van der Waals potential exist, the Lennard Jones 12-6 potential is commonly used.

$$V_{vdw} = \sum_{i=1} \sum_{j=1} (\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}) \tag{2.9}$$

where r_{ij} is the euclidean distance between atom i and j and A and B are atom pairwise parameters, each depending on the atom radius and well-depth parameter.

2.2.2 Molecular mechanics and dynamics

The hyper-surface of the potential energy of a macromolecule like a protein is very complex. Very often one is interested in determining the state of the system when the potential energy is low or minimal because these states correspond to stable configurations of the system. In order to *navigate* this potential energy surface towards local and global minima, so-called energy minimisation algorithms can be used. Methods such as *steepest descent* or *conjugate gradient* alter the system (change of atom positions) iteratively to identify pathways on the energy surface towards the minima. As the discussion of these methods is not central for the critical understanding of this thesis the reader is directed to corresponding literature [Leach, 2001].
Another approach called molecular dynamics (MD) make use of Newtons laws of motion relating the net force F directly to the mass and acceleration of a particle.

$$F_i = m_i a_i \tag{2.10}$$

$$\frac{dr_i(t)}{dt} = v_i \tag{2.11}$$

$$\frac{dv_i(t)}{dt} = \frac{F_i}{m_i} \tag{2.12}$$

where m is the mass, a the acceleration, F the force and v the velocity of a particle i. First simulations of molecular dynamics were performed in the 1960's [Alder and Wainwright, 1959] on 32 perfect spheres colliding with each other. To this date molecular dynamics have become a cornerstone in the study of protein dynamics and mechanistics, allowing the simulations of millions of atoms, increasingly long time-scales and its integration with experimental data to obtain information that, otherwise, would be inaccessible.
Although this thesis is not about molecular dynamics, on several occasions data used was produced using MD or related techniques. Thus principles are very shortly outlined here, but more focussed literature is recommended for the interested reader [Leach, 2001].

Integrating molecular motion

Simulating molecular motion of macromolecules involves the calculation of forces acting on all atoms in the system with all atoms in the system. Thus,

positions of atoms are highly interdependent and evolve together (or coupled) [Leach, 2001] which does not allow for an analytical solution of the equations of molecular motion. Instead numerical integration methods have to be employed.
Different numerical integration algorithms have been proposed in the past. For example an algorithm proposed by Verlet in 1967 uses the positions and accelerations of atoms at a time t and of a time $t - \delta t$ to calculate positions of atoms in time $t + \delta t$ [April, 2009]. However, the leap-frog algorithm is among the most widely used in macromolecular MD simulations.

$$r(t + \delta t) = r(t) + v(t + \frac{\delta t}{2}) \tag{2.13}$$

$$v(t + \frac{\delta t}{2}) = v(t - \frac{\delta t}{2}) + a(t)\delta t \tag{2.14}$$

The leap-frog algorithm calculates velocities at a time step $t + 1/2\delta t$ and these are used to calculate the new positions of atoms r at time step $t + \delta t$. Using these approximation, positions are calculated more precisely compared to the Verlet algorithm. The disadvantage is that velocities for time $t + \delta t$ are not directly available. However they can be derived using the following expression:

$$v(t) = \frac{1}{2}[v(t + \frac{\delta t}{2}) + v(t - \frac{\delta t}{2})] \tag{2.15}$$

Various other methods exist which are not covered here. A central aspect in all numerical integration techniques is the integration step. This step is crucially linked to the precision of the calculation. In the context of molecular dynamics simulations the integration step over time generally chosen has to be smaller than the fastest motion of the system. The covalent bond vibration is the fastest motion in a typical molecular system in classical mechanics.

Simulation conditions

Next to intra and intermolecular aspects in a system the conditions of a simulation can also affect the behaviour of it. Thus, prior to running a MD simulation it is required to fix an ensemble of these conditions:

- N: number of particles
- V: volume of the system
- T: temperature of the system
- P: pressure of the system
- E: energy of the system

Common ensembles are *(i)* the microcanonical ensemble or NVE, *(ii)* the isothermic-isobaric ensemble or NPT and *(iii)* the canonical ensemble or NVT. In MD simulation mostly NPT and NVT ensembles are employed.

Steered molecular dynamics

Classical MD simulations simulate freely evolving systems at equilibrium. However, lots of molecular processes occur at time-scales that are out of possible simulation times that can be achieved with classical MD. Thus, several approaches have been proposed to *"accelerate"* simulations. Here so-called steered molecular dynamics are shortly presented in the context of free energy calculations.
As the name indicates, steered MD or SMD directs a system along a particular reaction coordinate towards a wished end-state. A typical example related to drug design would be a MD simulation where a ligand is slowly pulled out of the binding site. Another interesting aspect in SMD is that the steering can happen either using a constant force or constant velocity. The second is particularly interesting because the force necessary to move from state A to B as well as the profile of forces on the trajectory A $\rightarrow$ B can be evaluated.
In 1997 Christopher Jarzynski derived a very powerful relation between the difference in free energy ΔG of sates A and B and the work W to go from A to B.

$$e^{-\Delta G/kT} = \langle e^{-W/kT} \rangle \tag{2.16}$$

where k is the Boltzmann constant and T the temperature of the system. The relation states that the average of the work necessary of taking a system from equilibrium state A to a non-equilibrium state B via all possible realisations approximately equals the difference in free energy between A and B. If the process of moving from state A to state B is done infinitely slow, ΔG equals W numerically.
SMD's jointly with the Jarzynski relation were used extensively within one work presented in this thesis. Practically, the Amber MD software was used to perform SMD simulations and extract the work (W). As the Jarzinsky relation is based on an average of the work, multiple SMD trajectories from different initial states (generally from 20 to 100) were produced and the Boltzmann average of the corresponding works calculated.

2.3 Protein pocket prediction

2.3.1 Introduction

The three dimensional structure of a protein is highly complex and dense. As shown on figure 2.8, the surface of a protein is irregular and contains protrusions, smaller and larger clefts as well as internal cavities. These pockets are of particular interest for target based drug discovery, that is to say, the discovery of small drug-like molecules fitting inside these pockets on particular proteins. The concave shape of a cavity usually allows binding of small molecules with higher affinities [Cheng et al., 2007, Englert et al., 2010]. It is mainly assumed that there is significant shape complementarity between a binding site and a small molecule inside the pocket and that this complementarity is responsible for a part of the affinity the ligand has for the receptor.

However, the concave shape notably allows the protein to modulate the behaviour of solvent and ligands on the protein surface. This modulation helps to ensure molecular recognition as well as right binding kinetics and thermodynamics. This line of thought is indirectly in accordance with results published by Kahraman and co-workers [Kahraman et al., 2007] showing that shape complementarity between small molecules and pockets is an often used assumption but turns out to be only partially true. Drug-like molecules are small and provide only a limited amount of putative interaction points. Thus, drug-like molecules usually interact through a series of weak directed (polar) as well as undirected hydrophobic interactions with the protein surface. If that interaction surface between the protein and the ligand was shallow, dissociation could be easily initiated via attacking solvent. Furthermore, the interaction surface between the ligand and the protein is maximised in such environments. Therefore a concave shape on the protein counterpart of a small molecule is required, which makes protein pocket prediction a relevant field for protein assessment in drug discovery.

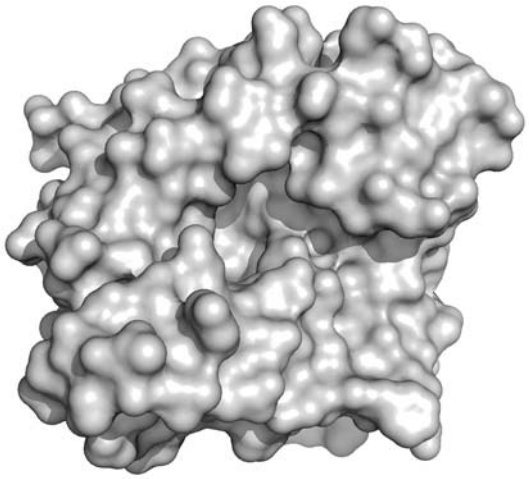


Figure 2.8: Surface representation of Heat shock protein 90 (rendered with PyMOL [DeLano, 2002])

2.3.2 Purpose

In order to understand the physico-chemical principles underlying protein ligand binding, a thorough analysis of the binding site and the ligand has to be undertaken. Protein cavities are complex 4 dimensional (space and time) entities difficult to *(i)* identify and *(ii)* analyse. While protein pocket identification is not a central part of this thesis itself, the main research results presented throughout this work are heavily based and connected to automated cavity prediction. Thus, principles, main players in the field and our own work is

shortly presented hereafter. A concluding discussion shows current pitfalls in this area of research and highlights progress to be made.

2.3.3 Pocket Identification Strategies

Repositories of structural information like the PDB can be used to gather a vast amounts of protein structures. As seen at least partially on figure 2.8, the shape of proteins is very complex. In order to study further properties of pockets on that surface, an automated way to detect these can be helpful. Pocket detection or identification means to automatically find the surface patch that corresponds to a known or putative binding site with sufficient accuracy. Fully automated identification of such binding sites is not trivial. During the last 2 decades, a vast amount of methods and programs have been proposed to address automated identification of protein ligand binding sites and it is still a very active field. Unfortunately, together with the increasing quantity of published methods, more and more approaches of how to validate these are considered and prediction assessment gets more diverse. While this might be a common evolution in scientific methods development it hinders strongly a thorough and objective comparison between all methods in the field.
In order to tackle computer based automated pocket predictions several approaches can be considered. All of them are shortly described hereafter highlighting only their principles.

Sequence based: These methods make use of the assumption that residues inside known pockets and active sites are conserved. This assumption is used for prediction of portions on the sequence corresponding to putative binding sites.

Energy based: Here a putative binding site is evaluated using energy calculations between probe atoms and the protein. The more favourable the potential interaction energy is, the more likely it is for the probe to be placed in the binding site of interest.

Geometry based: This type of methods takes advantage of the fact that cavities are, like the name indicates, concave shaped surface patches. Various, usually fast, geometrical principles can be applied to delimit a binding site.

Similarity based: Here known binding sites are used to derive structural patterns that describe these sites. These templates can then be used to compare a surface patch of a query protein and predict if this particular patch is a binding site similar to the template.

Mixed approaches and Meta-methods: The majority of existing methods in current literature use a mixture between mostly energy and geometry based approaches. Several other examples exist using various established methods for performing predictions and then consider the consensus of results as final pocket prediction.

Available methods & Software

Early works The program POCKET [Levitt and Banaszak, 1992] is among the first algorithms in the field efficiently using geometric properties of the protein surface to identify cavities. Using a regular grid probes are placed on each grid point and vectors are traced in the x, y and z dimension. Using this type of three dimensional scanning local buried parts on the protein surface can be identified. POCKET was the basis for various improvements. Figure 2.9 shows the principle of grid-based pocket detection algorithms schematically. Each grid point having more than one vector (or ray) cutting the protein below a given distance is considered to be in a buried portion (grey circles). The others are not (black x). Grid points close to known pocket grid points are often clustered together to form an ensemble of pocket grid points. The well known LIGSITE program [Hendlich et al., 1997] is an enhancement of the initial POCKET algorithm. Here not only 3 vectors are used for ray-tracing, but

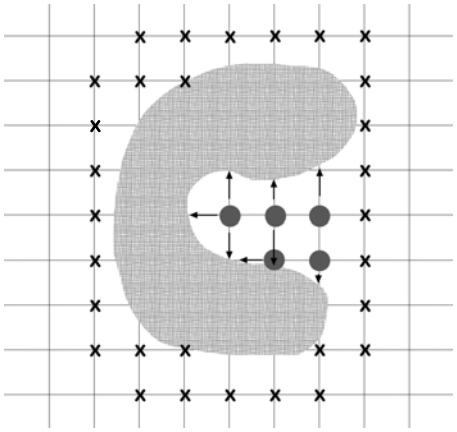


Figure 2.9: Scheme of a 2D cut of a protein with a superposed grid. Grid based pocket detection places probes on grid points and uses ray-tracing to delimit buried zones)

14 vectors are used. In subsequent improvements, LIGSITE was extended to account also for the Connolly surface [Connolly, 1983]. Instead of using atom coordinates, LIGSITE^{cs} performs ray-tracing on the accessible surface of the protein. A last extension was introduced to the algorithm by considering the conservation of residues in the identified binding site [Huang and Schroeder, 2006], thus introducing here sequence based evaluations next to pure geometrical techniques. While various approaches in the field make use of grids to detect surface depressions on proteins another type of geometry based approach is frequently encountered.
These techniques try to identify the biggest sphere one can fit to a surface portion without containing any atom of the protein. Several computational geometry means exist to perform this task. A seminal approach was published by Edelsbrunner and Liang making use of computational geometry principles until then not used for analysing biophysical properties of macromolecules [Edels-

brunner et al., 1995]. This so-called alpha shape theory was used to develop several approaches, like CAST [Liang et al., 1998] and subsequently CASTp [Dundas et al., 2006], SiteFinder (`http://www.chemcomp.com/journal/sitefind.htm`) and APROPOS [Peters et al., 1996]. Shown on figure 2.10, the alpha

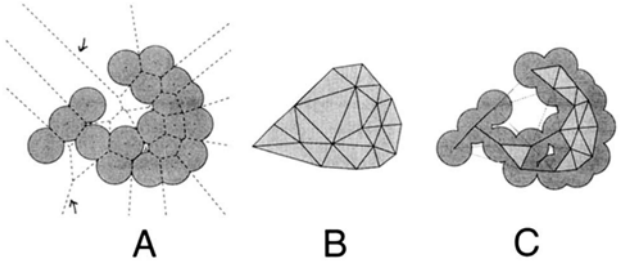


Figure 2.10: Figure and legend taken from [Edelsbrunner et al., 1995]: Illustration of concepts in alpha shape theory. **A:** A two-dimensional molecule consisting of of disks of uniform radii. The dashed lines show the Voronoi diagram of the molecule. Arrows indicate 2 of the 10 Voronoi edges that are completely outside the molecule. **B:** The convex hull of the atom centres in **A** (all shaded area) with Delaunay triangulation (triangles defined by dark lines). **C:** The alpha shape of the molecule in **A**. This alpha shape, or dual complex, consists of the light-shaded triangles, the dark line segments, and the atom centres. There are 10 shaded line segments corresponding to the 10 Voronoi edges that are completely outside the molecule. Any triangle with one or more shaded edges is an "empty triangle". A void formed by three empty triangles can be seen at the bottom center. It encloses a molecular cavity.

shape principles rely on three main principles. The protein is considered as a set of points in 3D space and this space is segmented using Voronoi tesselation [Voronoi, 1907]. Next, the convex hull is calculated to delimit the boundaries of the molecule and capture the main molecular shape, using the atom centres. Last, the atom centres are used to triangulate the hull[Delaunay, 1934] and isolate Voronoi edges outside the molecule.
Both geometrical approaches presented in this paragraph built the basis for upcoming improvements made in the field of pocket detection. While grid based approaches are easier to understand, they have a certain number of limitations. Using a grid to scan a protein surface results usually in algorithms with low computational performance. The results are dependent on the rotational and translational position of the protein and dependent on the grid-spacing used. Sphere fitting algorithms, like the one based on alpha spheres and subsequent developments show less sensitivity and are usually orders of magnitude faster than grid based algorithms.
Another method that should be mentioned in this section is called VOIDOO [Kleywegt and Jones, 1994]. Although, strictly speaking, it is not a pocket detection method, VOIDOO can be used to find internal cavities and invaginations on the protein surface. To do so and to measure accurately the cavity volumes, the program applies atom fattening techniques to close solvent accessible cavities and determine this way the volume. While this method has been frequently used for channel predictions, it is not suited to determine open

pockets on the protein surface.

Further developments in the field Among widely used programs in academics is PASS (Putative Active Sites with Spheres)[Brady and Stouten, 2000]. This algorithm coats the protein with several levels of spheres finally retaining sphere clusters that are sufficiently buried. SURFNET is also placing spheres between atom pairs to identify those not including any internal atoms [Laskowski, 1995].
In parallel, several new algorithms have been proposed. Notably, energy calculations to characterize the pockets made their entry in the field while computational resources became available to academic labs and force-fields more accepted by the community. All of these algorithms use principles similar to the widely used GRID program [Goodford, 1985], which places atom probes on an equally spaced grid around a protein and calculates interaction energies. Q-SiteFinder [Laurie and Jackson, 2005] developed in the Jackson lab in Leeds is among these energy based pocket identification algorithms. Similar in its principles, PocketFinder was proposed by Ruben Abagyans group [An et al., 2005] the same year. A contribution from 2008 [Morita et al., 2008] follows very similar ideas. All three approaches place methyl groups on a grid superposed to the protein and evaluate the interaction / Lennard Jones potential and use subsequent clustering and smoothing techniques to identify partially buried and "favourable" interaction sites. Probably the SiteMap program is the algorithm most tailored to the identification of drug binding sites [Halgren, 2007, 2009]. This software, marketed by Schrödinger LLC uses also energy calculations on an equally spaced grid. Geometric delimitation of the pocket is done using principles similar to the first geometry based algorithms in the field, POCKET and LIGSITE. Subsequently, energy calculations are mainly used to assess and characterize the pocket, assign it a so-called SiteScore and rank all cavities according to this score.
Pure geometry based algorithms using grids have also been further considered compared to initial approaches like POCKET and LIGSITE [Venkatachalam et al., 2003, Weisel et al., 2007]. However, given the multitude of available algorithms in this field another type of approach has seen the day. In a protocol called Metapocket [Huang, 2009], Huangs group gather results from different cavity detection programs, namely PASS [Brady and Stouten, 2000], Q-SiteFinder [Laurie and Jackson, 2005], SURFNET [Laskowski, 1995], fpocket [Le Guilloux et al., 2009], GHECOM [Kawabata and Go, 2007], ConCavity [Capra et al., 2009] and POCASA [Yu et al., 2010]. Metapocket takes only the top 3 cavities from each prediction method and clusters spatially close cavities together.
The scope of this thesis is not to compare and describe all algorithms in the field. Cavity prediction is a very active field and still today several new methods see the day. Unfortunately there is no critical review encompassing the whole field available. Papers coming closest to this, but not assessing these methods were recently published [Pérot et al., 2010, Ghersi and Sanchez, 2011]. The reasons for the lack of a thorough review of all methodologies are complex. They are tentatively discussed in section 2.3.4.

Methodological differences & consequences As seen previously, various different ways exist to identify cavities on complex surfaces such as a protein surface. Each implementation and algorithm has its advantages and disadvantages and the final choice of how to implement a method or which method to use depends on the purpose of the work. A relatively important number of existing methods use grids superimposed to protein structures. These methods tend to be computationally more expensive than alternative geometry based approaches (for instance). Importantly, energy based approaches make an important protein preparation step necessary. Atom types and force-field parameters have to be assigned accurately to use these methods, connected to an even higher computational cost. Grid based methods furthermore render cavity prediction results dependent on the position of the protein within the grid. Despite these facts, grid based and especially energy based methods have advantages. An energy based evaluation of the protein surface can give first and important insights into the characteristics of the binding site. Often underestimated, but grid based methodologies are usually easier to understand than their geometry-based counterparts. Energy-grid-based methodologies can thus be deemed an excellent choice for punctual and accurate pocket detection needs that involve further characterization of the cavities.
Other geometry based approaches (alpha shapes for example) are computationally very efficient and do not suffer theoretically from rotational sensitivity of results. However, these methods usually do not characterise the pockets that were identified. Usually very simple descriptors encompassing size and enclosure are extracted and used to rank one binding site versus another.

2.3.4 Current pitfalls

Pocket identification is an active field of research for nearly 20 years now. During this time more than 30 different algorithms, programs and improvements of existing tools have been published.

Datasets: Like in other disciplines, for example small molecule docking, major problems persist in the objective evaluation of such methods using a common benchmark and conditions to compare prediction performance. Unlike in docking, no common benchmarks like the DUD [Huang and Schroeder, 2006] or the Astex Diverse Set [Hartshorn et al., 2007] have been thoroughly established or consequently used within the domain. While Nayal & Honig [Nayal and Honig, 2006] proposed a comprehensive study on 99 proteins, other studies like [Huang and Schroeder, 2006] provided data-sets and methods comparison which were subsequently used more often in literature [Weisel et al., 2007, Le Guilloux et al., 2009, Zhu and Pisabarro, 2011].

Precision: Another central question when evaluating these algorithms is: When is a cavity correctly identified and with what precision? While some methods like Q-SiteFinder [Laurie and Jackson, 2005] focus on evaluating thoroughly the prediction performance, others like PASS [Brady and Stouten, 2000], LIGSITE^{csc} [Huang and Schroeder, 2006], PocketPicker [Weisel et al., 2007] or

PocketFinder [An et al., 2005] use more permissive criteria to define what a correctly identified binding site and the precision of the prediction is.

The identification paradox: Pocket detection algorithms are meant to identify putative binding sites. One protein can have more than one cavity on its surface. Subsequently the identified sites are usually scored and ranked between each other to identify the cavity that resembles most to *the binding site*. Such scoring and ranking is often based on training a score with a set of known protein ligand complexes.
In order to evaluate the performance of a pocket prediction method one detects all cavities on a set of different proteins with known binding sites. A successful pocket prediction method is deemed to rank a significant amount of known binding sites from different proteins among the first few positions. However, identification and ranking are completely independent tasks that should also be evaluated separately.
For example, all recent pocket prediction algorithms report high success rates (90%) for ranking *cognate binding sites* from a benchmark data-set among the first 3 ranks. Unfortunately, pure identification success is rarely reported, meaning how many binding sites among a set of known binding sites are identified.
Jayaram and Singh show that the pocket identification algorithm they present is able to identify all cognate binding sites. This is also known to be the case for fpocket (not published data) [Jayaram and Singh, 2011, Le Guilloux et al., 2009]. It is very likely that a substantial amount of programs allow the identification of all binding sites. Thus, it should be acknowledged that modern pocket identification algorithms are excellent, unlike stated in the concluding remarks of a recent review in the field [Pérot et al., 2010].

Significantly more work and especially focus is however necessary on classifying types of cavities, an entirely different problem. Hormone binding sites from nuclear hormone receptors are very distinct from galactose binding sites for instance. Furthermore, a protein can have multiple cavities on its surface. Among these only one might act as active site (for an enzyme for example), but others could act as allosteric sites, peptide binding sites etc. In conclusion, the main focus of the field should move from cavity identification to characterization of very precise types of cavities.
For example, Liu and Altmann proposed FEATURE, a method to identify specifically Ca^{2+} binding sites in disordered regions. Another active field is trying to predict the druggability of cavities, thus characterising them as being able to bind drug-like molecules [Fauman et al., 2011]. Other approaches focus on sugar binding sites [Taroni et al., 2000, Shionyu-Mitsuyama et al., 2003], nucleotide binding sites [Kono and Sarai, 1999] or peptide binding sites [Petsalaki et al., 2009], but no pocket prediction program is able to classify cavities according to what they are likely to bind.

Availability of methods: Despite the important number of methods in the field, no clear trend in publishing new methods as freely accessible tools can

be observed. Cavity detection is connected to financially interesting drug discovery, thus various methods are only commercially available. Others are only available as public web-servers, making them appealing for academic groups, but unusable for pharmaceutical industry. Other methods are freely available, but only binaries are provided. Thus, no further extensions of project specific improvements can be implemented if necessary. Finally, only a fraction of methods is free, provide source code and a web-server for running the program. A tentative review of accessibility of various software is available in [Ghersi and Sanchez, 2011].

Proteins move: Another aspect usually not considered in such studies is that proteins are dynamic and pockets can move. The problem is that a crystal structure of a protein provides only a static snapshot. Integration of protein motion in binding site prediction and estimation of pocket plasticity are of utmost importance but mostly unmet by currently available approaches. A dedicated section 2.4 explains current developments in this field as this is an issue of high importance linked to a part of the work presented in this thesis.

2.3.5 The fpocket project

Prior to my thesis, Vincent le Guilloux and myself, started the fpocket project in the context of the MSc in Bioinformatics of the University Paris Diderot. Next, the principles behind fpocket are explained.
Edelsbrunner and Liang established the use of the alpha shape theory for cavity detection and fpocket uses a very similar principle. Using Voronoi tesselation performed by Qhull [Barber et al., 1996], the 3D space defined by protein atoms is segmented into so-called Voronoi cells, defined by Voronoi edges. Crossing Voronoi edges form Voronoi vertices, as shown on figure 2.11. Such Voronoi ver-

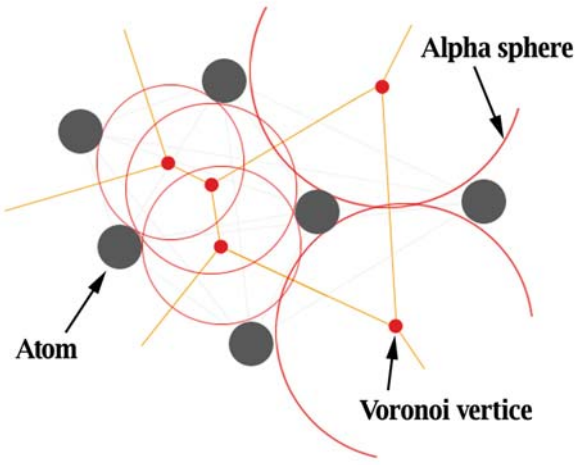


Figure 2.11: Schematic representation of Voronoi tesselation on a protein.

tices have a property that used by fpocket. Onto each Voronoi vertice a sphere can be placed that contacts exactly 4 atoms of the protein without any internal atoms in the sphere volume. Such a contact sphere is called an alpha sphere. Depending on the shape of the surface portion an alpha sphere is contacting the

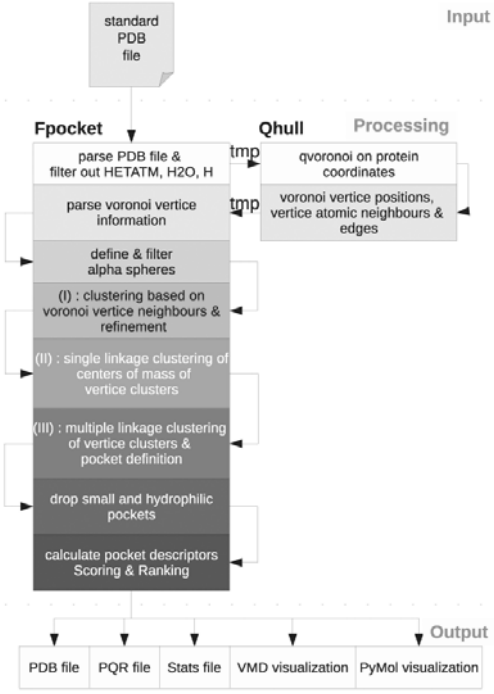


Figure 2.12: Overall fpocket workflow for cavity detection

radius of the latter will vary from very small to infinite. The algorithm applies a simple filtering step retaining only alpha spheres big enough not to fill interstitial space between atoms on the protein interior, but also small enough to only encompass areas of a certain degree of enclosure. Next several clustering steps are performed to gather together alpha spheres situated in close neighbourhood with respect to each other. Finally, such an alpha sphere cluster is retained if a certain number and types of alpha spheres compose the cluster. A pocket is thus formed by such a cluster, describing the void space lined by the atoms contacted by all alpha spheres of the cluster.
The clustering and filtering procedures can be controlled by the user via command-line flags and are explained in detail in the fpocket manual and paper [Le Guilloux et al., 2009]. Various pocket descriptors are calculated by fpocket and a score is attributed to each pocket, allowing pocket ranking. I has been shown that fpocket achieves very good prediction performance on predicting protein ligand binding sites in reference benchmark sets in the literature [Le Guilloux et al., 2009]. Furthermore, the program is among the fastest in the field, open-source and free.

2.4 Transient pocket and channel prediction

2.4.1 Small molecule binding sites

The majority of pocket detection methods for small molecule binding sites concentrates on pointing surface locations on fixed, representative structures of proteins. The reasons for the lack of methods accounting for protein flexibility are multiple. Accounting for protein flexibility is complex and experimental data is rare. Thus one has to use computational usually very costly techniques like molecular dynamics (MD), to get a partial insight into the protein plasticity. Surprisingly, only one approach exists to detect transient pockets on molecular conformational ensembles created using MD. This protocol named EPOS^{BP} was proposed by Eyrisch and Helms and was assessed on the detection of transient cavities on protein-protein interfaces [Eyrisch and Helms, 2007]. Using the cavity detection algorithm PASS [Brady and Stouten, 2000] the protocol systematically defines cavities on each frame of the MD trajectory and clusters distinct cavities together when they significantly overlap using the contacted residues of the protein as reference.

Novelty: This study tackled several aspects in the field of drug discovery. First the authors consider that even very short living cavities can be interesting for drug discovery purposes. Second, they consider targeting protein protein interfaces, notoriously difficult due to their shallowness, detecting situations in time when these cavities deepen, to putatively accommodate a small molecule. Interestingly, several currently known interfacial binders act as stabilizer of a particular protein conformation unable to interact with its protein counter-part.

Limitations: Transient cavity appearance is considered to be very short lived in this study. While there is still an ongoing discussion on if ligand protein association implicating protein motion is mainly due to an induced fit effect or conformational sampling, the authors claim the second possibility appropriate. One main limitation in the current version of EPOS^{BP} is the fact that pockets identified on different time-frames but in the same regions can be considered as separated pockets with a discrete pocket ID. This fact is hindering to gain insights into continuous pocket and channel plasticity. This limitation becomes obvious while trying to visualise transient cavities as an ensemble. EPOS^{BP} was only accessible via a public web-server after the publication of the work, however it was recently released as standalone binary. Unfortunately, neither for PASS [Brady and Stouten, 2000] nor for EPOS^{BP} source code is available. Usage and especially visualization and interpretation is limited. Only few properties can be assessed using the approach.

2.4.2 Ligand migration channels

Significantly more work has been done in a related field to pocket prediction. In addition to surface depressions, proteins can also contain migration channels or very small internal cavities. These channels and cavities are of importance

for functioning of some types of proteins, like myoglobin. Here, the size of a cavity or channel is able to allow migration of a diatomic ligand at most and thus cavities are an order of magnitude smaller than typical drug binding sites. Although initially not developed for this purpose VOIDOO [Kleywegt and Jones, 1994] allows internal cavity and volume calculations and doing so on a set of conformations from MD trajectories allows to assess plasticity of internal channels and cavities. However, the use of VOIDOO for this purpose is rather time consuming and not straightforward.

The CAVER family

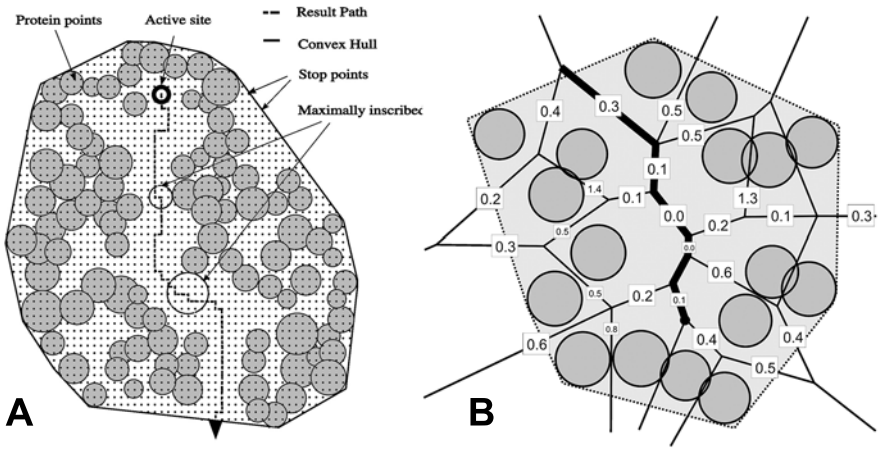


Figure 2.13: Channel search methods implemented in CAVER (**A**) and MOLE (**B**). Figure adapted from [Petrek et al., 2006, 2007]

CAVER [Petrek et al., 2006, Beneš et al., 2010] is a PyMOL plugin and standalone program that allows internal channel detection on MD trajectories. CAVER was recently improved to a software called MOLE using computational geometry principles instead of grid-based calculations [Petrek et al., 2007]. As schematically shown on figure 2.13 A, CAVER superimposes a grid to the protein structure. A cost value is associated to each grid-point. This value is related to the maximum sphere one could place at this point, just touching the protein surface. Then a starting position is defined (active site). Next, a modified version of the Dijkstra algorithm for detection of the shortest path [Dijkstra, 1959] is used to find optimal exit pathways on the grid coordinates. MOLE uses very similar principles, but the path detection is guided by Voronoi edges obtained from Voronoi tesselation of the protein structure as shown on figure 2.13 B. Subsequently again the Dijkstra algorithm is used to identify cost-effective paths. A very similar approach to MOLE has been followed in a recent contribution by Yaffe and coworkers proposing a method called MolAxis for detection of channels from the interior of the protein to the bulk solvent [Yaffe et al., 2008]. All these previously cited methods claim to allow calculation of channels on single crystal structures and molecular dynamics tra-

jectories. This is only partially true, as channels have to be identified on single static structures. Once identified, these can then be analysed using several structures of a molecular dynamics trajectory. There is a fundamental difference between observing transient channels on an ensemble of conformations of a macromolecule and observing a putative channel on a static structure and then track it.

Novelty: All algorithms of this category have similar principles, aims and outcomes. Main novelties from these methods are intuitive visualisation of cavities (CAVER & MOLE) or more rapid calculations (MolAxis). Furthermore, channel dimensions can be measured on molecular dynamic trajectories.

Limitations: Here again, these methods suffer from practical limitations. Most importantly, all are purpose built to detect internal channels from a starting position to the outside of the protein. As discussed previously, these methods allow channel extraction on static structures and deriving transientness of appearance of channels from conformational ensembles is insufficiently covered.

Implicit ligand sampling

Another very distinct technique frequently used for studying transient gas migration pathways is implicit ligand sampling, or ILS. Unlike methods described before, ILS is a technique that allows identification of transient channels on molecular dynamics trajectories [Cohen et al., 2008]. Prior to analysis, all snapshots of the MD trajectory are aligned (structurally superimposed) to a reference structure. Next an equally spaced grid is placed onto the protein. Similarly to the GRID program [Goodford, 1985], a probe molecule (O_2 for example) is placed on each grid point and the interaction energy between the probe and the protein is calculated. Here the interaction energy is only approximated with the Lennard Jones potential. Next, formula 2.17 is used to derive the free energy of placing a gas molecule on each grid point. This energy is implicit, as no explicit gas particle was simulated during the MD trajectory, but it is considered that these channels can transiently open without the presence of a diatomic ligand.

$$G_{implicit}(r) = -k_B T ln \sum_{n=1}^{N} \sum_{k=1}^{C} \frac{e^{-\frac{\Delta E_{n,k}(r)}{k_B T}}}{NC} \tag{2.17}$$

$G_{implicit}$ is the free energy of placing the ligand on position r. k_B is the Boltzmann constant and T the temperature. N is the number of frames used from the molecular dynamics trajectory and C the number of conformations and rotations considered for the gas particle. Finally $\Delta E_{m,k}$ is the interaction energy calculated at position (r) on a given frame (n) with a given rotation of the gas particle.
Applying this calculation on a whole MD trajectory, free energy maps can be

derived allowing easy visualisation of transient migration pathways, inferred implicitly.

Novelty: ILS is the first approach in the field to use sound physical principles to address exploratory transient gas migration pathway prediction. The result of ILS is an energy map, that can be visualised using popular molecular viewers, like VMD [Humphrey et al., 1996], PyMOL [DeLano, 2002] or Chimera [Huang et al., 1996].

Limitations: Energies derived using ILS are implicit and it is inferred that the gas particle has no influence on transient openings of channels. Second, only van der Waals interaction energies are calculated in a reasonable amount of time to allow analysis of MD trajectories. To track specific channels or calculate putative migration pathways, subsequent and independent analysis tools have to be used. Last, no evaluation of the influence of the structural superimposition, necessary for this method, on the results is performed.

DyME

More recently, a MATLAB protocol called DyME has been introduced [Lin and Song, 2011] allowing also detection of migration pathways that transiently appear on MD trajectories. This procedure performs a Voronoi tessellation on the protein on each snapshot of a MD trajectory that has been previously superimposed. Voronoi vertices with small minimal boundary spheres (clearance) are discarded. All remaining vertices are transformed to a spanning tree. Spanning trees for all frames of the MD trajectory are overlapped and conserved cavities are clustered together. Portal regions, connecting internal channels to the outside are identified. In a final super graph conserved cavities are connected via maximum clearance channels observed at least once during the MD simulation.

Novelty: DyME is relatively fast compared to other commonly used tools. Furthermore, it allows to produce an easy to visualise and understand channel network with associated channel radii (clearances).

Limitations: The article describing DyME is missing important details to discuss fully limitations of the approach. Like for other methods, no assessment of the necessary structural superimposition has been done. Furthermore, it is likely that migration pathways identified as maximum clearance pathways are inferred, as the maximum clearance radius can be rather low.
Last, the method is not available as standalone tool to the scientific community.

2.5 Pocket characterization

2.5.1 Definition of druggability

The term "druggability" is used for designating very different concepts in the area of genome analysis and drug discovery. In the work presented in this the-

sis, the druggability is associated to a binding site, and therefore to a protein. A druggable binding site is a pocket that can bind a molecule similar to orally available marketed drugs. This definition does not take into account any consideration about pharmacokinetics, the mechanism of action or toxicity.
The definition of druggability is necessarily linked to the notion of drug-likeness. Here again debate is ongoing on the scope of drug-likeness.

2.5.2 Controversy on the term *"Druggability"*

The word "Druggability" is a neologism that appeared first in abstracts of scientific papers in 1999 [Labischinski and Johannsen, 1999]. However, the "Druggable Genome" paper by Hopkins and Groom [Hopkins and Groom, 2002] made the first comprehensive review on chemical tractability using the catchy word "druggable" in 2002. Since that review, the usage of the term became widespread among different areas in drug discovery. Figure 2.14 shows the number of scientific papers referenced in Pubmed containing the word "druggable", "drugable" or the corresponding nouns in the title or the abstract. Despite the

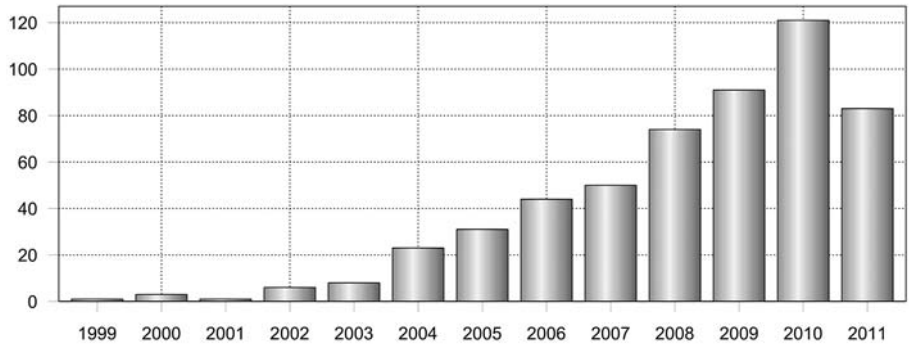


Figure 2.14: Number of publications with the word "druggable" or "druggability" in the title or abstract.

more widespread usage of this neologism in today's literature, the exact scope of the term remains fuzzy and varies substantially from paper to paper. The current wikipedia entry on druggability highlights the ambiguity of the definition : *"Druggability is a term used in drug discovery to describe the suitability of a protein or protein complex to be targeted by a drug or drug-like molecule, in a way that this interaction will alter the proteins function and correct disease causing behavior. The term is typically used for designating tractability by a small molecular weight drug, although druggability can be achieved using biotherapeutics such as monoclonal antibody. Druggability is now an ubiquitously used term that is referred to in different contexts, but always meaning the suitability of a target for drug discovery"*
From a purely structure based point of view, the usage of this term is too generic. The actual structural features responsible for the interaction between a drug and the protein only constitute a part of all features required for a

protein to be druggable. Structure based techniques do not assess the disease relation of proteins and if the binding of a small molecule has an impact on the function of the protein. As stated by Robert Sheridan and co-workers at Merck : *"The term "druggable" implies too much. Getting to an actual drug involves many hurdles that are almost impossible to predict in advance and involve properties of small molecules as well as the target. For example, there is nothing in those approaches that determines whether the active site in the target is different enough from related targets that selectivity is possible. It is also not clear that there will be the desired therapeutic effect in vivo even if we can find drug-like molecules to bind to the target protein."* [Sheridan et al., 2010]. Thus, the term of "chemical tractability" should be used, renamed by Sheridan et al to "bindability". These terms encompass the definition of druggability, but without the restrictions on unpredictable properties from structural data. The reason why finally the term druggability was used throughout this work is mostly due to two factors : (i) if we train a method on known druggable proteins, we have explicit knowledge about the druggability and tractability in the training set, (ii) for sake of shortness, druggability is a less cumbersome terminology than chemical tractability. If the exact definition and, most importantly, restrictions are given in the context of using the word druggability we deemed it appropriate using it instead of chemical tractability. This fact is further underlined by several authors in the field deeming the usage of "druggability" appropriate, knowing the underlying approximations and assumptions.

2.5.3 Definition of *"non-druggable"*

Defining a target druggable using historical results that can be found in literature is a manageable task. On the contrary, to define a target non-druggable we have to accept a substantial amount of approximations and suppositions. The first data-sets differentiating druggable from non druggable proteins in a structure based context were provided by Hajduk et al. [Hajduk et al., 2005a] from Abbott and later by Alan Cheng and co-workers at Pfizer [Cheng et al., 2007]. While the first deemed proteins non-druggable if the NMR fragment screening hit was low on these targets, the second declared a protein non druggable if despite substantial efforts (several years of research) drug discovery projects fail. The limitation of both approximations is that they are dependent on the (i) NMR fragment library and experimental conditions and (ii) on the fact that systematically all possible techniques have been tried to identify drug-like molecules on these targets.

2.5.4 Prediction of druggability

Despite the controversy on what druggability is, several groups tried to develop methods to predict if a given gene and associated proteins are tractable or not. Within this thesis I'll focus especially on structure based methodologies that use 3D protein structures and properties of putative binding sites to assess if they could bind small drug-like molecules.
To develop methods predicting druggability, reference data has to be found to

describe both druggable and non druggable binding sites. While historically most binding site identification methods focused on the identification of drug binding sites, they did not specifically assess the nature of the small molecules binding in these pockets. Probably the dataset derived by Emanuele Perola and co-workers [Perola et al., 2004] comes closest to the first set of druggable pockets. This dataset was subsequently used by Nayal and Honig to extract an extensive set of pocket descriptors for these cavities [Nayal and Honig, 2006]. However, no consistent definition of non druggable cavities has been given till the pioneering work of Hajduk, Huth and Fesik [Hajduk et al., 2005a].

Abbott Laboratories (Hajduk et al)

In this pioneering methods paper Hajduk and colleagues follow a very simple and appealing idea. The more hits can be achieved using experimental NMR based screening, the more the target is deemed druggable. This assumption was then used to characterize a set of proteins as druggable and non druggable (hit rate at 0). Next the *ActiveSite_Search* flood-fill algorithm from Insight II was used to identify binding sites computationally on the protein surface. Various characteristics of the identified pocket were derived and used to built a regression based model predicting experimental NMR hit rates. The predictivity of the different descriptors was assessed and their final contribution to the score is symbolized by a + if positive, and - if negative. The NMR hit rate can be predicted using the total surface area of the pocket (+), the polar contact area (-), the apolar contact area (+), the first principal moment (-), the third principal moment (+) and the pocket compactness (+). The pocket compactness is defined by the ratio between the pocket volume and surface area. The principal moments tend to capture the shape of the pocket. The final model predicts experimental screening hit rates with a leave-one-out Q^2 of 0.56.

Novelty: This study was the first to derive a statistical model to predict experimental screening hit rates, assumed herein to be correlated with druggability. Furthermore, it is the first method using automated pocket prediction to perform this task. Potential applications of this method are shown in the review published by the same group the same year [Hajduk et al., 2005b].

Limitations: Despite the certain impact this study had on the field and the novelty, the work has several limitations especially regarding reproducibility of the results presented in the study. The authors use a set of proteins among which a substantial amount of inaccessible in-house structure are listed. Furthermore, for non-profit research groups, the usage of software like Insight II (discontinued, now Discovery Studio) for pocket detection and descriptor extraction is connected to licensing costs.

Pfizer (Cheng et al)

A significant impact in the field had the study performed by Alan Cheng and co-workers [Cheng et al., 2007]. The reasons for the popularity of this work

are diverse. First it was published in a high impact journal (Nature Biotechnology), gathering a broad readership compared to more focused journals in the field. Second, the work presented a newly derived and this time publicly available data set for knowledge based druggable and also non-druggable proteins. Last, the model that was derived in this work is very simple and easy to understand, so are the underlying principles and easy ideas tend to propagate faster in science.
A total of 63 crystal structures corresponding to 27 different pharmaceutical targets from the PDB were selected from literature as basis for a new data-set in the druggability prediction field. A target was defined non druggable if despite efforts in the pharmaceutical industry, no known drug has seen the day. Out of these 27 targets, 4 were deemed non druggable. Furthermore, Cheng and co-workers distinguished druggable proteins from pro-drug binding proteins, creating thus a third category containing 6 targets. The remaining 17 targets were defined as druggable.
All crystal structures used in this study were co-crystallized with a small molecule inside the known binding site. A mix of automatic binding site detection (using SiteFinder implemented in MOE) and the environment of the cognate ligand was used to define the extent of the pocket on the protein structure. Computational geometry algorithms have been used to describe the binding site and derive descriptors, like the curvature of the pocket and surface areas [Liang et al., 1998, Coleman et al., 2005].
Using these descriptors a model was created meaning to predict the maximum affinity a drug-like molecule could ideally have for the given binding site. The model called $\mathrm{MAP_{POD}}$ is shown in equation 2.18

$$\Delta G \approx -\gamma(r) A_{apolar}^{target} \frac{A_{druglike}^{target}}{A_{total}^{target}} + C \tag{2.18}$$

Here $\gamma(r)$ is a term dependent on the curvature of the binding site. $A_{druglike}^{target}$ is a constant value considered to be 300 Å^2. A_{target}^{apolar} corresponds to the apolar accessible surface area of the pocket and A_{total}^{target} is its total accessible surface area. Basically, the model relates hydrophobic enclosed to maximal affinity that can be achieved. As shown in the results of this work, the model is able to distinguish between druggable proteins and prodrug-binding and non-druggable proteins.

Novelty: Several problems have been addressed by this paper. Of importance for methods development is the provided data-set with publicly available structures. The paper shows furthermore, that without complex regression and a few assumptions a predictive model can be produced.

Limitations: Despite the novelty of the data-set proposed, it is still very small and the choice of crystal structure for a given target seems arbitrary. The definition of the binding site is semi-automatized and thus needs human intervention and the presence of a small molecule. This fact renders the method

hardly usable for screens of structural databases for unoccupied druggable binding sites. Indeed a purely automated pocket detection could be used to delimit the pocket extent, but the curvature calculation algorithms used to derive the radius of the pocket are certainly very elegant, but unfortunately very sensitive to changes in the definition of the pocket (non-published reimplementation of the method in-house and communication with Ryan G Coleman). These limitations render the method hardly usable as a screening tool.

Schrödinger (Halgren)

Thomas Halgren from the New York based Schrödinger LLC published the so-called Dscore in 2009 [Halgren, 2009]. The previously established set of structures by Cheng et al was used to fit a regression based model using cavity descriptors from SiteMap [Halgren, 2007], the binding site detection program sold by Schrödinger. The model is shown in formula 2.19

$$Dscore = 0.094\sqrt{n} + 0.60e - 0.324p \tag{2.19}$$

Here n is the number of site points for the cavity (grid voxels), e is the degree of enclosure for the site and p a hydrophilicity score computed by SiteMap. Using only 3 descriptors, SiteMap and Dscore are able to predict druggability by reproducing results published by Cheng et al. Furthermore, Dscore appears to be able to distinguish between non druggable proteins and prodrug binding proteins.

Novelty: The principal novelty is that the derived Dscore is able to spot the three categories defined by Cheng et al, (i) druggable, (ii) prodrug binding and (iii) non druggable. Else, the paper follows a similar idea as the work proposed by the group from Abbott, linking automatic pocket detection to an automated score using pocket properties. A definite plus for Dscore is that it's available for purchase from Schrödinger and thus potential customers can test and evaluate the score themselves.

Limitations: The derived score is based on the Cheng data set, that is fairly small. The multiplicative factors of the descriptors of the model have been derived using the very same data, so no thorough learning / validation process has been performed. Although SiteMap in combination with Dscore is the first ready to use tool in literature, cost of the software might be prohibitive for punctual use within non profit research labs. Last, the use of grid based energy calculations renders the pocket prediction and descriptor extraction using SiteMap relatively slow and the protein preparation cumbersome. While accuracy can be important for punctual predictions, a large scale screen of structural database would be difficult to perform.

Merck (Sheridan et al)

A fundamentally different approach was followed by Robert Sheridan and coworkers from Merck contributing the last method published so far in this area of

research [Sheridan et al., 2010]. Starting from a purely ligand based viewpoint. All pockets containing a drug-like ligand in the PDB are considered tractable. For all pockets three basic properties are calculated: (i) the volume, (ii) the buriedness and (iii) the hydrophobicity. These three properties are used to create a three dimensional pocket space. Using this space so called drug like densities (DLID) are derived. The DLID basically assesses the number of neighbouring binding sites containing a drug like ligand a cavity has in the three dimensional pocket space. The rational presented in this work suggests the more pockets in the immediate neighbourhood bind drug like ligands, the more the particular pocket is likely to be chemically tractable.
In a next step Sheridan et al. built a simple regression model (mDLID) using these three pocket characteristics (formula 2.20) :

$$mDLID(v, b, h) = -8.70 + 1.71log(v) + 3.94(b) + 2.27(h) \tag{2.20}$$

where v is the volume of the pocket, b the buriedness and h the hydrophobicity. mDLID appears to be well correlated to the previously calculated DLID and can thus be used to approximate it. Although no training was done on druggable binding sites explicitely, DLID is shown to correlate with previous results published by Cheng et al. and Halgren.

Novelty: The approach chosen by this group was very distinct from other contributions made to the field so far. Based on a ligand drug like point of view chemical tractability and not druggability was assessed. This way it is possible to analyse several thousands of structures instead of doing "manual" training on a hundred structures. The manuscript highlights several problems related to large scale structural database analysis and also regarding the definition and limitations of the term druggability. While these are not a fundamental novelty, it should still be highlighted that they have been critically assessed in this paper in the contrary to several other contributions in the field.

Limitations: The misuse of druggability according to Sheridan et al led to the use of the term "bindability" or chemical tractability meaning the capability of pockets to bind drug-like molecules. In this particular work, the use of the term druggability is indeed inappropriate. Thus, the outcome is not directly comparable to previous work performed on druggability prediction. In any way this study incorporates the strength of an observed interaction in the crystal structure and thus even pockets accommodating weak binders that cannot be optimized further could be potentially classified tractable. While this is true for using the terminology tractability or bindability, it isn't druggability.
More importantly, the authors define a pocket space using three pocket descriptors and use this space to derive pocket neighbourhoods to calculate the DLID. The subsequent relation of the DLID to mDLID is thus predefined and a correlation not that astonishing as claimed by the authors. Last and probably the most critical point concerns the choice of the descriptors. This choice is not thoroughly discussed in this work, but the set of descriptors are of great significance as they are used to represent the pocket space in the later analysis. Thus herein the authors consider that the descriptors chosen are best suited to

efficiently separate tractable from non tractable pockets in the pocket space. While this choice can be considered at best as an educated guess (especially regarding existing literature in the field), a clear rational for descriptor selection is missing.

Parallels between all methods

It is very impressive to see that all methods consider the same type of descriptors without exception. Of utmost importance appears to be the hydrophobicity of the cavity. All characteristics that are generally used and tracked down by all these models are well resumed in Halgrens work [Halgren, 2009]. Among

Category	Characteristics
undruggable	→ very strongly hydrophilic; little or no hydrophobic character → requires covalent binding → very small or very shallow
difficult	sufficient hydrophobicity for administration as prodrug; less hydrophobic as a typical cavity
easy	reasonable size & enclosure (buriedness, curvature) and hydrophobicity with unexceptional hydrophilicity

Table 2.1: Typical characteristics of druggable, non druggable and difficult pockets. Adapted from Halgren [Halgren, 2009].

hydrophobicity, relative buriedness and the size play a role directly or indirectly in all models. In consequence all druggable pockets are ideal when they are completely hydrophobic and very buried and of reasonable size. Interestingly no model considers the importance of polar atoms in the binding site.

Other methods in the field

The previously presented methods all calculate characteristics of pockets and try to use these to predict druggability. As the models are all fairly simple and based on usually easy to calculate descriptors they tend to be fast. Another category of methods does not consider only the protein for determining druggability but ligand protein interactions.
Inspired by fragment screening these computational methods assess whether small molecular fragments or solvent molecules bind into specific locations on the protein surface. CS-MAP (for Computational Solvent MAPping) published by the Vajda group is one of these methods[Landon et al., 2007]. The protocol intends to identify interaction hotspots by systematically placing 14 different small molecular fragments on the surface of the protein and assess the interaction energy of these. The method is shown to be able to distinguish regions binding drug-like molecules from other regions binding peptidomimetics for example. Furthermore apparent correlation between NMR hit rates and the

energy of clustered fragments in a pocket exist. This is shown comparing CS-MAP results to the public structures in the training set published by Hajduk et al[Hajduk et al., 2005a].
Within our group another approach has been developed using molecular dynamics (MD) trajectories of proteins with a solvent mixture of isopropanol and water [Seco et al., 2009]. The occupancy of isopropanol on the protein surface is then related to a maximum affinity estimate that a drug-like molecule could have for this site. The applicability of this method is shown on several characterised and known systems like MDM2, PTP1B, elastase and more. As maximum affinities are derived from dynamically occurring contacts (based on molecular mechanics force-fields) no prior training has to be performed so the method is not tailored to a given training set. Furthermore it is able to identify protein protein interaction sites and other shallow surface patches. Thus, this methodology named MDMix is till today probably the most unbiased and accurate measure of putative fragment interaction patterns that can be observed on the protein surface.
A last category of methods uses systematic molecular docking of small fragments into binding sites to assess their suitability to bind drug-like molecules. Analogue to experimental NMR hit screening the approach published Huang and Jacobson relates virtual screening hit rates to druggability[Huang and Jacobson, 2010].
This second category of methods has an inherent advantage of not being dependent on extensive training or pocket properties of predefined characteristics. Thus they tend to be more generic and should be the methods to choose when one particular target has to be evaluated. However, if the aim of druggability prediction is not a single target validation, but exploratory putative target discovery, then these techniques are far too complex and slow to be applied on structural databases.

Importance of druggability prediction

Several papers and reviews highlight the importance of druggability predictions for pharmaceutical industry usually in introductory paragraphs or conclusions[Egner and Hillig, 2008, Fauman et al., 2011]. Also seen the list of herein presented methods it shows that main pharmaceutical companies like Abbott, Pfizer and Merck work on these issues and the fact that software companies like Schrödinger take the effort to implement their own druggability prediction score shows that there can be potential interest in the industry. Despite this apparent interest, other pharmaceutical groups appear uninterested towards further target assessment (personal communication with representatives from Johnson & Johnson, Actelion and others). In most scientific communications in the field especially from pharmaceutical industry one can observe that the main aim of predicting druggability is to validate a particular target. The use of such methods for large scale screens to discover new putative targets is still in its beginnings in the field.

Bibliography

B J Alder and T E Wainwright. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics*, 31(2):459, 1959.

Jianghong An, Maxim Totrov, and Ruben Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & cellular proteomics : MCP*, 4(6):752–61, June 2005.

Jean M Standard April. The Verlet Algorithm for Molecular Dynamics Simulations. *Time*, (6):4–6, 2009.

C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.

Petr Beneš, Chovancová Eva, Kozl\'iková Barbora, Pavelka Anton\'in, Strnad Ond\vrej, Brezovský Jan, Šustr Vilém, Klva\vna Martin, Szabó Tibor, Gora Artur, Zamborský Matúš, and Biedermannová Lada. CAVER 2.1, 2010.

G P Brady and P F Stouten. Fast prediction and visualization of protein binding pockets with PASS. *Journal of computer-aided molecular design*, 14(4):383–401, May 2000.

John A Capra, Roman A Laskowski, Janet M Thornton, Mona Singh, and Thomas A Funkhouser. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS computational biology*, 5(12):e1000585, December 2009.

Alan C Cheng, Ryan G Coleman, Kathleen T Smyth, Qing Cao, Patricia Soulard, Daniel R Caffrey, Anna C Salzberg, and Enoch S Huang. Structure-based maximal affinity model predicts small-molecule druggability. *Nature biotechnology*, 25(1): 71–5, January 2007.

Jordi Cohen, Kenneth W Olsen, and Klaus Schulten. Finding gas migration pathways in proteins using implicit ligand sampling. *Methods in enzymology*, 437:439–57, January 2008.

Ryan G Coleman, Michael A Burr, Diane L Souvaine, and Alan C Cheng. An intuitive approach to measuring protein surface curvature. *Proteins*, 61(4):1068–74, December 2005.

Michael L Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558, 1983.

Andrew M Davis, Simon J Teague, and Gerard J Kleywegt. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angewandte Chemie (International ed. in English)*, 42(24):2718–36, June 2003.

Warren L DeLano. The PyMOL Molecular Graphics System, 2002.

B Delaunay. Sur la sphère vide. *Bulletin of Academy of Sciences of the USSR*, 7(6): 793–800, 1934.

E W Dijkstra. A note on two problems in connection with graphs. *Numerische Mathematik*, 1(269-270):269–271, 1959.

Joe Dundas, Zheng Ouyang, Jeffery Tseng, Andrew Binkowski, Yaron Turpaz, and Jie Liang. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic acids research*, 34(Web Server issue):W116–8, July 2006.

H. Edelsbrunner, M. Facello, Ping Fu, and Jie Liang. Measuring proteins and voids in proteins. page 256, January 1995.

Ursula Egner and Roman C Hillig. A structural biology view of target drugability. *Expert Opinion on Drug Discovery*, 3(4):391–401, April 2008.

L Englert, A Biela, M Zayed, A Heine, D Hangauer, and G Klebe. Displacement of disordered water molecules from hydrophobic pocket creates enthalpic signature: binding of phosphonamidate to the S_1'-pocket of thermolysin. *Biochimica et biophysica acta*, 1800(11):1192–202, November 2010.

Eran Eyal, Sergey Gerzon, Vladimir Potapov, Marvin Edelman, and Vladimir Sobolev. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *Journal of molecular biology*, 351(2):431–42, August 2005.

Susanne Eyrisch and Volkhard Helms. Transient pockets on protein surfaces involved in protein-protein interaction. *Journal of medicinal chemistry*, 50(15):3457–64, July 2007.

Eric B Fauman, Brajesh K Rai, and Enoch S Huang. Structure-based druggability assessment-identifying suitable targets for small molecule therapeutics. *Current opinion in chemical biology*, 15(4):468–463, June 2011.

Dario Ghersi and Roberto Sanchez. Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures. *Journal of structural and functional genomics*, 12(2):109–17, July 2011.

P J Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry*, 28(7):849–57, July 1985.

Philip J Hajduk, Jeffrey R Huth, and Stephen W Fesik. Druggability indices for protein targets derived from NMR-based screening data. *Journal of medicinal chemistry*, 48(7):2518–25, April 2005a.

Philip J Hajduk, Jeffrey R Huth, and Christin Tse. Predicting protein druggability. *Drug discovery today*, 10(23-24):1675–82, December 2005b.

Thomas a Halgren. Identifying and characterizing binding sites and assessing druggability. *Journal of chemical information and modeling*, 49(2):377–89, February 2009.

Tom Halgren. New method for fast and accurate binding-site identification and analysis. *Chemical biology & drug design*, 69(2):146–8, February 2007.

Michael J Hartshorn, Marcel L Verdonk, Gianni Chessari, Suzanne C Brewerton, Wijnand T M Mooij, Paul N Mortenson, and Christopher W Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of medicinal chemistry*, 50(4):726–41, March 2007.

M Hendlich, F Rippmann, and G Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of molecular graphics & modelling*, 15(6):359–63, 389, December 1997.

Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature reviews. Drug discovery*, 1(9):727–30, September 2002.

Bingding Huang. MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omics : a journal of integrative biology*, 13(4):325–30, August 2009.

Bingding Huang and Michael Schroeder. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC structural biology*, 6 (1):19, January 2006.

C C Huang, G S Couch, E F Pettersen, and T E Ferrin. Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components. *Pacific Symposium on Biocomputing*, 1:724, 1996.

Niu Huang and Matthew P Jacobson. Binding-Site Assessment by Virtual Fragment Screening. *PLoS ONE*, 5(4):10, 2010.

W Humphrey, A Dalke, and K Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–8, 27–8, February 1996.

B Jayaram and Tanya Singh. AADS- An automated active site identification, docking and scoring protocol for protein targets based on physico-chemical descriptors. *Journal of chemical information and modeling*, page null, August 2011.

Abdullah Kahraman, Richard J Morris, Roman A Laskowski, and Janet M Thornton. Shape variation in protein binding pockets and their ligands. *Journal of molecular biology*, 368(1):283–301, April 2007.

Takeshi Kawabata and Nobuhiro Go. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins*, 68 (2):516–29, August 2007.

G J Kleywegt and T A Jones. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta crystallographica. Section D, Biological crystallography*, 50(Pt 2):178–85, March 1994.

H Kono and A Sarai. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, 35(1):114–31, April 1999.

Harald Labischinski and Lars Johannsen. Cell wall targets in methicillin-resistant staphylococci. *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy*, 2(5):319–325, October 1999.

Melissa R Landon, David R Lancia, Jessamin Yu, Spencer C Thiel, and Sandor Vajda. Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. *Journal of medicinal chemistry*, 50(6):1231–40, March 2007.

R A Laskowski. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, 13(5):323–30, 307–8, October 1995.

Alasdair T R Laurie and Richard M Jackson. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)*, 21(9):1908–16, May 2005.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10:168, January 2009.

A R Leach. *Molecular Modelling: Principles and Applications*, volume 2nd. Prentice Hall, 2001. ISBN 0582382106.

B Lee and F M Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–400, February 1971.

D G Levitt and L J Banaszak. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of molecular graphics*, 10(4):229–34, December 1992.

J Liang, H Edelsbrunner, and C Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein science : a publication of the Protein Society*, 7(9):1884–97, September 1998.

Tu-Liang Lin and Guang Song. Efficient mapping of ligand migration channel networks in dynamic proteins. *Proteins*, 79(8):2475–90, August 2011.

Harvey Lodish, Arnold Berk, Paul Matsudaira, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, S. Lawrence Zipursky, and James Darnell. *Molecular Cell Biology*. W.H. Freeman & Company, 2004.

Mizuki Morita, Shugo Nakamura, and Kentaro Shimizu. Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins*, 73(2):468–79, November 2008.

Murad Nayal and Barry Honig. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, 63(4):892–906, June 2006.

Emanuele Perola, W Patrick Walters, and Paul S Charifson. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins*, 56(2):235–49, August 2004.

Stéphanie Pérot, Olivier Sperandio, Maria A Miteva, Anne-Claude Camproux, and Bruno O Villoutreix. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug discovery today*, 15(15-16):656–67, August 2010.

K P Peters, J Fauck, and C Frömmel. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *Journal of molecular biology*, 256(1):201–13, February 1996.

Martin Petrek, Michal Otyepka, Pavel Banás, Pavlína Kosinová, Jaroslav Koca, and Jirí Damborský. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC bioinformatics*, 7(1):316, January 2006.

Martin Petrek, Pavlína Kosinová, Jaroslav Koca, and Michal Otyepka. MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure (London, England : 1993)*, 15(11):1357–63, November 2007.

Evangelia Petsalaki, Alexander Stark, Eduardo García-Urdiales, and Robert B Russell. Accurate prediction of peptide binding sites on protein surfaces. *PLoS computational biology*, 5(3):e1000335, March 2009.

Gale Rhodes. *Crystallography Made Crystal Clear, Third Edition: A Guide for Users of Macromolecular Models (Complementary Science)*. Academic Press, 2006. ISBN 0125870736.

Jesus Seco, F Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *Journal of medicinal chemistry*, 52(8):2363–71, April 2009.

Robert P Sheridan, Vladimir N Maiorov, M Katharine Holloway, Wendy D Cornell, and Ying-Duo Gao. Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *Journal of chemical information and modeling*, 50(11): 2029–40, November 2010.

C. Shionyu-Mitsuyama, T. Shirai, H. Ishida, and T. Yamane. An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins. *Protein Engineering Design and Selection*, 16(7):467–478, July 2003.

A H Stouthamer. A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie van Leeuwenhoek*, 39(3):545–65, January 1973.

C. Taroni, S. Jones, and J. M. Thornton. Analysis and prediction of carbohydrate binding sites. *Protein Engineering Design and Selection*, 13(2):89–98, February 2000.

Peter Tsvetkov, Nina Reuven, and Yosef Shaul. The nanny model for IDPs. *Nature chemical biology*, 5(11):778–81, November 2009.

C M Venkatachalam, X Jiang, T Oldfield, and M Waldman. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of molecular graphics modelling*, 21(4):289–307, 2003.

G Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques, 1907.

Martin Weisel, Ewgenij Proschak, and Gisbert Schneider. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central journal*, 1(1):7, January 2007.

Eitan Yaffe, Dan Fishelovitch, Haim J Wolfson, Dan Halperin, and Ruth Nussinov. MolAxis: efficient and accurate identification of channels in macromolecules. *Proteins*, 73(1):72–86, October 2008.

Jian Yu, Yong Zhou, Isao Tanaka, and Min Yao. Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics (Oxford, England)*, 26(1):46–52, January 2010.

Hongbo Zhu and M Teresa Pisabarro. MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics (Oxford, England)*, 27(3):351–8, March 2011.

Chapter 3

Results and Discussion

3.1 Prediction of protein druggability

3.1.1 Introduction

Since the publication of the ground-breaking article by Hopkins & Groom [Hopkins and Groom, 2002] describing the druggable genome lots of discussion emerged on the concept of druggability and ways to predict it using structural and sequence information available in the public and private domain.

The concept and the term of druggability vary substantially from one study to the other. Throughout this work, the definition given by Hopkins & Groom is used. Thus, a protein is deemed druggable if it has at least one pocket on its surface capable of binding molecules having drug like properties regardless of the molecular mechanism of action. A more thorough discussion on the scope of this term can be found in section 2.5.2.
One of the main reasons of the high impact of the *druggable genome* paper was the estimation that there might be *only* around 3000 genes encoding for putatively druggable proteins out of the supposedly 22000 human genes [Hesman Saey, 2010]. This estimation created lots of speculation on the actual number of tractable proteins in the human but also viral, bacterial, fungal and other genomes. *"Surprisingly, for an industry that spends in excess of US$50 billion on research and development each year, there is a lack of knowledge of the set of molecular targets that the modern pharmacopoeia acts on. If we are to develop predictive methods to identify potential new drug targets, it is essential that we establish with confidence the number, characteristics and biological diversity of targets of approved drugs."* [Overington et al., 2006]. This introduction in John Overingtons review entitled "How many drug targets are there?" highlights an apparent paradox linked to this ongoing discussion in literature.
The results published by Overington et al. reproduced in figure 3.1 show that there is currently a clear tendency in pharmaceutical research to target well characterized target families, like G-protein coupled receptors, nuclear hormone receptors and ion channels, accounting alone for more than 50% of all targets identified in that survey. Also serine/threonine and tyrosine kinases and other proteins are among well studied targets, but till today with lower outcome on actual drugs on the market.

Structure based druggability Druggability became a hot topic for a part of the pharmaceutical industry accompanying steady discussion about reasons for high attrition rates together with the observation of a decrease in the number of new targets for new chemical entities (NCE's) discovered. Indeed it would be invaluable for emerging drug discovery projects to be able to assess, very early, if a given target is actually able to accommodate a drug-like molecule and second if it is disease-related. While the second is usually a pre-requisite for biologists to validate a protein as putative therapeutical target, the first is not. Furthermore, current drug discovery and target discovery pipelines don't sieve systematically for new putative targets using structural information avail-

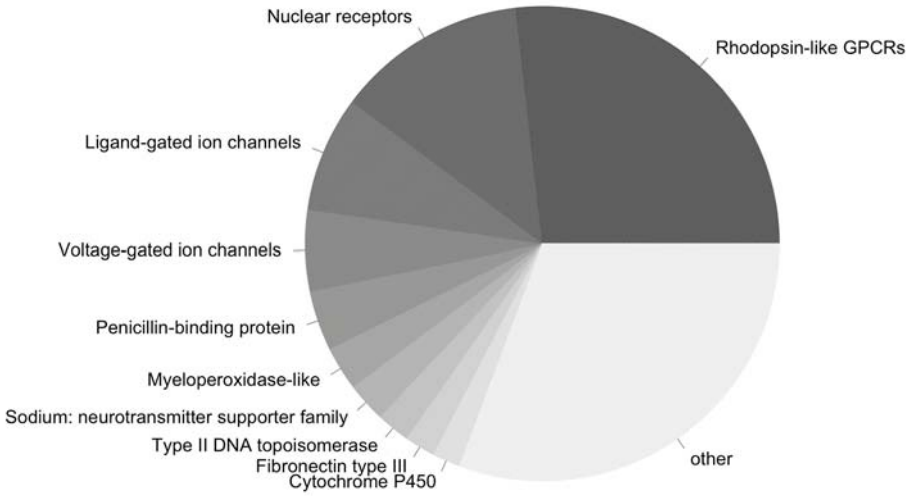


Figure 3.1: The family share as a percentage of all FDA-approved drugs for the top ten families. Adapted from Overington et al.

able in public databases like the PDB or in-house data banks.
During the last years a few notable methods emerged trying to use the 3D structure of the protein and the known location of the binding site to study the properties of the latter and assess if it would be able to bind a drug like molecule. An initiating contribution has been done by Hajduk et al. describing a first regression model to predict the likelihood of binding small molecular fragments into binding sites [Hajduk et al., 2005]. Another very simple model has been described by Cheng et al. in 2007 [Cheng et al., 2007]. This model related size, shape and relative hydrophobicity of the binding site to druggability. Furthermore this paper proposed the first public dataset for evaluation of structure based druggabilty descriptors. Last, Tom Halgren from Schrödinger LLC implemented a druggability prediction into the already existing pocket detection software called SiteMap [Halgren, 2007, 2009]. The resulting Dscore is based on SiteMap descriptors of the binding site like enclosure, size and hydrophilicity. All methods in the field are explained in more detail in section 2.5.4.
Despite the publication of these methods, none has been used to efficiently sieve structural databases for yet uncharacterized druggable binding sites. Second, they have been validated only on a restricted dataset. Third, none of them is freely available and ready to be used. Last, most of these methods inherently assume a druggable binding site to be hydrophobic. While hydrophobicity is a frequent observation in drug binding sites and a good predictor of druggability, a purely hydrophobic pocket would be unable to yield specificity towards drug like molecule.

3.1.2 Objectives

Focusing on the development of a new structure based druggability prediction method, this work intends to address four main issues in the field.

Dataset availability and extension Current datasets in the field are very small regarding the number of publicly available proteins they include and number of conformational states per protein. Based on the already available data, this work intends to propose a larger dataset of protein binding sites known to recruit drug like molecules and pockets known to be undruggable. Unlike previous publications, the dataset should be publicly available. [Hajduk et al., 2005] Furthermore, not only a hand picked ensemble of structures per protein should be used, as done by Cheng et al [Cheng et al., 2007] and Halgren [Halgren, 2009], but available structures of a protein should be systematically used if they fulfil necessary quality requirements.

Coupled to automated cavity detection In contrast with several previous works in the field the method developed here should use automatic 3D pocket detection algorithms to delineate the protein binding site.

Large scale applicability The focus of this work is somehow further restricted, as the method is intended to be applicable on huge structure databases, like the PDB. Thus the reliability and accuracy of the pocket identification has to be weighted against a high throughput usage. Next, the druggability prediction needs to be robust towards variations of the structure of the protein as well as variability occurring during cavity detection. A part from the necessary scientific accuracy and reliability another important objective in high-throughput applications is technical feasibility. This includes notably reasonable calculation time and usage of computer resources.

Investigate the role of polar atoms in drug binding sites Interestingly, the role of polar atoms in drug binding sites is completely neglected in druggability predictions. One objective of the work presented here is to highlight the importance and presence of polar atoms within druggable binding sites.

Open access to our methods and data A central point of my work is related to open access. Unfortunately it is very common in research related to drug discovery to publish methods that are then further commercialized. Often these methods are out of reach for potentially interested researchers in academia or in small companies. Another recurrent problem is that they usually can not be tailored to a specific context or use inside a research group because of disclosed source code of underlying methods.
Publishing results and methods in open accessible source code allows other researchers to honestly evaluate and use the research results immediately, also increasing the impact of the published results.

3.1.3 Results

Creating and sharing a novel data set

Training knowledge based scoring functions to predict properties like "druggability" requires a substantial set of diverse structural data. Previous datasets available to the community were either partially disclosed or representative structures had been chosen in a semi objective manner. Furthermore, as the creation of such a set is mainly based on lengthy literature analysis and experience, only few proteins were included (27 for the only publicly available data set).
Prior to the development of a new druggability assessment I established a new reference data-set. This new data-set uses the existing data provided by Cheng and co-workers [Cheng et al., 2007] as well as the public part of the data provided by Abbott [Hajduk et al., 2005]. Next, supplementary data was added via a semi-automatic protocol.
A list of marketed and orally administered drugs [Vieth et al., 2004] has been cross referenced with PubMed [Bolton et al., 2008] and the PDB [Berman et al., 2000] to identify if the PDB contains co-crystallized structures of the drug and the protein. The retrieved protein-ligand complexes were further validated using DrugBank [Wishart et al., 2006] to verify that the protein actually corresponds to the intended target of the drug. Next, the resulting complexes were manually classified into three categories already pre-defined by Cheng and co-workers: (i) druggable, (ii) non-druggable, (iii) prodrug binding. For the first time, also apo structures have been added to the data-set.

	NRDD		DD		Cheng	Hajduk
druggability class	**holo**	**holo**	**apo**	**total**	**holo**	**holo**
druggable	45	773	146	919	17(43)	35
nondruggable	20	75	9	84	4 (10)	37
prodrug	5	60	7	67	6 (10)	
total	**70**	**908**	**162**	**1070**	**27 (63)**	**72**

Table 3.1: Composition of the DD dataset. NRDD: non redundant data-set (one structure per protein). For the Cheng data-set the number of proteins are listed and number of structures in brackets.

A list of 1070 crystal structures from 70 different proteins was derived. As the assembly of such a set of proteins is intended to be helpful for the community different ways can be envisioned for the distribution of the data. Usually a list of structures is published in a scientific communication and the data-set either becomes a benchmark in the community or other contributors consider establishing a new data-set. To keep the data set creation and shaping accessible at all times to the community a scientifico-sociological experiment was connected to this work. I created a collaboratorive web-platform named "Druggable Cavity Directory" (DCD) that allows researchers from all over the

world to anonymously download and browse the data. The DCD contains assigned druggabilities on a scale of 1 to 10 (10 being druggable) connected to a PDB code and the hetero-atom residue name to spot the ligand in the binding site (defining the binding site). In addition, researchers are allowed to register as contributor and/or validator to further add new data to the existing data-set. Thus it is possible for the community to continuously evolve the data, discuss the pertinence of certain assignments and use it to retrain / validate new methods and models. The druggable cavity directory is accessible on: http://fpocket.sourceforge.net/dcd

Are polar atoms needed in drug binding sites?

This new and more extensive data-set was then put to use to analyse structural properties of druggable binding sites versus non druggable binding sites. It should be noted that all properties were extracted from the pocket definition from cavities automatically identified using fpocket. These pockets have to overlap significantly with the known binding site (mutual overlap criterion has to be 1, c.f. [Le Guilloux et al., 2009]) While previous studies highlighted the importance of the hydrophobicity, size and enclosure of the binding site, none of these [Hajduk et al., 2005, Cheng et al., 2007, Halgren, 2009] considers investigating the role of polar atoms in the binding site. This results in the debatable assumption that druggable binding sites are necessarily very greasy and buried. Polar interactions are known to be important as anchoring points playing a central role in pharmacophoric models of binding sites [Hoffmann, 2006]. Thus, during the analysis of available pocket descriptors for druggability prediction special attention was also paid on factors that can influence drug recognition by the protein.
Very simple surface properties of binding sites from the DD have been calculated. First the polar accessible surface area was compared to the total surface area of the pocket. Previous studies of surface properties by Cheng [Cheng et al., 2007] already suggested that for druggable pockets the accessible surface area is only to 20% to 40% polar. As shown on figure 3.2 this trend is also seen within the DD. On the contrary, non druggable binding sites tend to be more polar (40-60%). From the analysis of around 220.000 cavities identified by fpocket on single protein chains in the PDB around 50% of their surface area is defined by polar atoms (±16%). Again this result highlights the im-

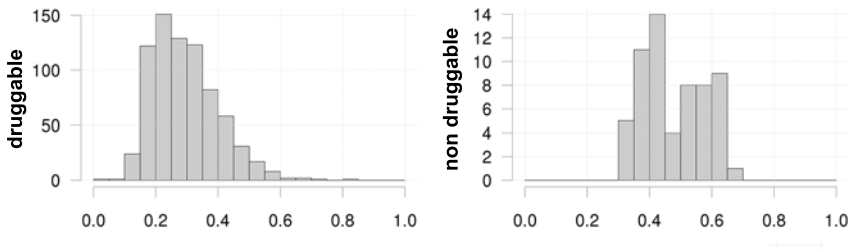


Figure 3.2: Accessible surface ratio between polar and total area for druggable and non druggable cavities.

portant hydrophobicity of druggable binding sites. Another characteristic that has been observed is that in druggable cavities, 70% of all polar atoms have an accessible surface area of 10Å^2 versus 50% in non druggable cavities.
Hydrogen bonds can be up to 1.2kcal/mol stronger in hydrophobic environments [Gao et al., 2009]. To study the protrusion of polar atoms into the hydrophobic cavities another property has been calculated on all known druggable and non druggable binding sites. In order to know to what extent polar atoms make themselves available, the polar and apolar accessible surface areas have been calculated. This step was repeated increasing the van der Waals radius r_i of the pocket atoms from r_i+1.4 $\AA$ (for water) to r_i+2.2 $\AA$. By doing so, the cavity gets gradually filled up and surface area decreases due to the concave shape of the pocket. This process called atom fattening is also used in several other computational approaches like VOIDOO [Kleywegt and Jones, 1994], but not for the same purpose.
While tracking polar and apolar surface areas during atom fattening, different scenarios can be imagined like shown in figure 3.3 A, B and C. Here three

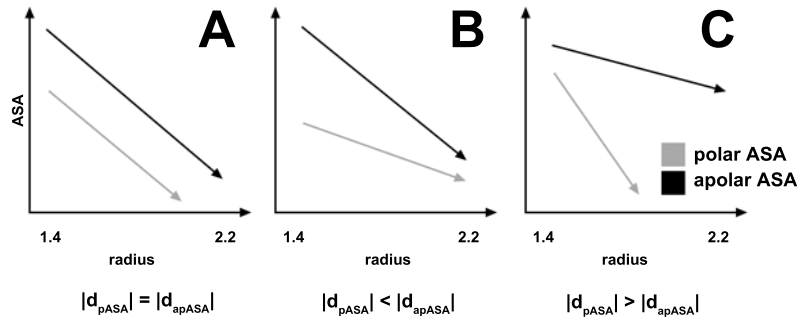


Figure 3.3: Schematic representation of polar versus apolar surface areas (ASA) upon atom fattening in different binding sites (radius axis)

cases have been considered. Given that the cavity is by definition globally concave, the surface area is decreasing or, at most, stable during atom fattening. On figure 3.3 A a case is represented where the polar (grey curve) and apolar (black curve) loose accessible surface area upon atom fattening at the same pace. Thus the derivative of the curves d_{pASA} and d_{apASA} are equal. Next, the relative position of the apolar versus the polar ASA curve can vary as already shown on figure 3.2. In figure 3.3 B the polar surface area decreases less fast compared to the apolar surface area. This indicates that polar atoms are more exposed than the apolar surface area. The contrary is shown in C, where apolar atoms are more exposed than polar atoms.
Now the relative decrease in polar and apolar surface area has been analysed on druggable and non druggable cavities and it was found that druggable cavities systematically follow a profile like the one shown on figure 3.3 B. This is more clearly shown on figure 3.4, where the ratio between d_{apASA} and d_{pASA} (called concavity profile ratio) is shown for druggable and non druggable cavities. On this figure we can see a clear shift of this ratio towards values bigger than 1 for

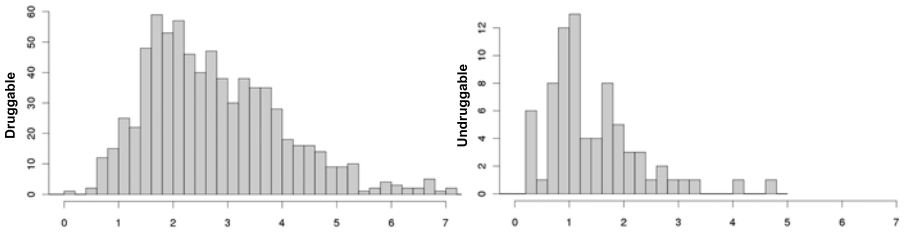


Figure 3.4: Ratio between apolar and polar slopes of ASA curves upon atom fattening

druggable proteins. This means that the polar surface area, although less important (c.f. figure 3.2) is more exposed compared to the apolar surface area. On the contrary for non druggable proteins, most pockets have a concavity profile ratio close to 1 or below 1.
While it is known that druggable binding sites are more apolar than other pockets, the protrusion of poorly exposed polar atoms in these is significantly different from non druggable binding sites. These specific polar atoms can thus be seen as hallmark of druggable binding sites. This might infer that their protrusion can play a fundamental role augmenting their visibility to putative interaction partners. In section 3.2 a subsequent work is presented intending to provide insights into why these characteristics are observed and especially how the local environment around polar atoms can affect the interaction between the pocket and a putative ligand molecule.

Coupling to an automated pocket prediction method

The main objective of the work presented in this section was the development of a new druggability prediction method. To fulfil all requirements previously defined an automatic cavity prediction algorithm had to be used to detect, define and characterise all putative binding sites on a protein 3D structure. During the last two decades a vast amount of such algorithms have been proposed and they are summarized in section 2.3 and reviewed in [Pérot et al., 2010]. Taken all algorithms and programs in this field only two programs could be considered for this project: ligsite^{csc} [Hendlich et al., 1997] and fpocket [Le Guilloux et al., 2009]. Both algorithms have good pocket prediction accuracy [Le Guilloux et al., 2009, Schmidtke et al., 2010], are open source and reasonably fast to allow large scale pocket predictions without the need of high performance computing facilities.
As fpocket was developed as a project for the MSc Bioinformatics Paris by Vincent le Guilloux and myself and published in the beginning of my thesis the choice between ligsite^{csc} and fpocket was facilitated by the fact that I simply knew all details, advantages and disadvantages of fpocket. Next to an initial small scale evaluation of fpocket [Le Guilloux et al., 2009], we further validated fpocket in a large scale comparison with 3 commercial pocket identification algorithms, namely ICMPocketFinder [An et al., 2005], SiteMap citepHalgren2007 and SiteFinder included in the Molecular Operating Environment (MOE). In

this study (annexed to this thesis) [Schmidtke et al., 2010], fpocket was found to be the fastest algorithm together with SiteFinder while producing the most accurate binding site predictions. SiteFinder and SiteMap had higher success rates in binding site identifications, but often with a cost of producing unreasonably big binding sites.
Once a cavity is identified by fpocket two scenarios could be considered: *(i)* The cavity location known via fpocket is used to re-analyse the pocket and derive new descriptors tailored for druggability predictions or *(ii)* the cavity descriptors already extracted by fpocket during pocket detection are used for druggability predictions. For this project the second scenario was considered as it is more suitable for large scale applicability. Furthermore it allows easy adaptation of fpocket to include a new druggability score and its subsequent dissemination.

A model predicting druggability

All descriptors available in fpocket were extracted for all cavities from a non redundant version of the DD (druggability data-set defined for this study) hereafter named NRDS, acronym for non redundant data-set. The terminological distinction between druggability and non druggability is very strict and purely binary, although from a scientific point of view this distinction could be mapped on a continuous scale. A mathematical model having a close to binary behaviour while being continuous is the logistic model. As the logistic function is sigmoid, a prediction can vary quickly from being 0 (non druggable) to 1 (druggable), having an inflexion point at 0.5, which would be the druggability threshold. While this is a wanted behaviour it can be problematic knowing that pocket definitions can vary from one structure to another for the same protein. Thus a second layer of logistic models was incorporated for the final druggability scoring model to behave like a two step voting model. First a logistic model is separately constructed for each descriptor. Then the results from the predictions of these models are used in a final logistic model weighting previous "votes" to a final score. The final model is shown in equation 3.1. Using this modelling layout, the R statistical software was used to test vari-

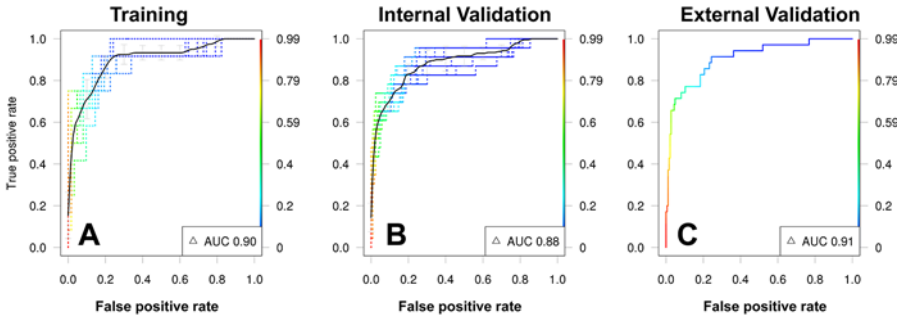


Figure 3.5: ROC curves for internal learning (A) / validation (B) and external validation (C) of the druggability score.

ous combinations of descriptors for their *(i)* predictive power for druggability and *(ii)* stability upon cross validation and external validation. To this end the NRDS was split in two parts constituting a training set and an external validation set. Next the training set was used to perform a 10 fold bootstrap cross-validation on half of its data. A set of three descriptors have been isolated via this procedure and receiver operator characteristics (ROC) for various modelling steps are shown on figure 3.5. Finally, the other half of the NRDS was used to validate the model. It should be noted that previously assessed descriptors regarding polar accessible surface area and the protrusion of polar atoms were also considered in this assessment. However, unfortunately these descriptors were found to be too variable regarding different pocket definitions and were thus discarded as candidate descriptors for a high-throughput model predicting druggability.

$$ds(z) = \frac{e^{-z}}{1 + e^{-z}} \tag{3.1}$$

$$z = \beta_0 + \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3 \tag{3.2}$$

$$f_x(d_x) = \frac{e^{-\beta_{x,0} + \beta_{x,1} d_x}}{1 + e^{-\beta_{x,0} + \beta_{x,1} d_x}} \tag{3.3}$$

descriptor	**coefficient**	**mean**[a]	$\frac{stdev}{mean}$ [b]
intercept	β_0	-6.238	-0.095
mean local hydrophobic density (norm)	β_1	4.592	0.154
hydrophobicity score	β_2	5.717	0.170
polarity score (norm)	β_3	3.985	0.459
intercept	$\beta_{1,0}$	-5.141	-0.170
mean local hydrophobic density (norm)	$\beta_{1,1}$	6.579	0.173
intercept	$\beta_{2,0}$	-2.669	-0.168
hydrophobicity score	$\beta_{2,1}$	0.056	0.216
intercept	$\beta_{3,0}$	-2.445	-0.238
polarity score (norm)	$\beta_{3,1}$	2.762	0.330

Table 3.2: Coefficients for the final druggability score after 10 fold boostrap cross validation.

The learning and validation process allowed us to derive a set of parameters for each descriptor of the model 3.1. These parameters are summarised in the table 3.2. The three fpocket descriptors retained for the final druggability model are described in more detail in the fpocket paper [Le Guilloux et al., 2009] and hereafter:

1. The Mean local hydrophobic density is a numerical value assessing the amount of local hydrophobic sub-pockets in a binding site and their relative packing. Shortly, the number of overlapping apolar alpha spheres in a binding site is calculated and normalized by the count of apolar

alpha spheres in the binding site. In the druggability model, the normalized mean local hydrophobic density is used. This normalization was performed versus all other pockets on the same protein structure.

2. The hydrophobicity score is a measure of the propensity of hydrophobic residues in the binding site using a hydrophobicity scale of amino-acids [Monera et al., 1995]. Here the average hydrophobicity score is calculated for the whole pocket.

3. The polarity score is a score classifying the polarity of amino acids lining the binding site. Two categories of polarities (0 or 1) were considered. This was originally published on a web resource at the university of Angers (France), which is not accessible anymore. The polarity score values for each amino-acid are now publicly accessible in the fpocket source code (file aa.c here : `http://fpocket.sourceforge.net/programmers_guide/aa_8c-source.html`). The final polarity score used as descriptor is the average of all polarity scores for each amino acid lining the binding site. Each residue is evaluated only once.

As shown on figure 3.5 stable predictions with good enrichment in druggable cavities are obtained during internal learning and validation as well as external validation.
Interestingly all descriptors have a global positive contribution to the final druggability score. As highlighted in a recent review [Fauman et al., 2011], this makes this particular druggability score the first model to incorporate a polar term as positive contribution for predicting protein druggability and is a counter example to all structure based druggability estimation models that exist in the field.

Variability upon different crystal structures

A central point of predictions in high throughput applicable methods is their evaluation versus variability. Even though several methods have been published, their predictions were usually validated using only one single crystal structure, although hundreds might exist in the PDB. The DD presented previously was thus subsequently used to measure the variability of druggability prediction of a given target using fpocket automated pocket detection and scoring using the druggability score developed throughout this work. Figure 3.6 resumes predictions on all structures in the DD for each protein in the DD for all three categories considered (druggable, non druggable and prodrug binding or difficult). Analysing these results we can observe that certain categories yield very stable and correct predictions, like buried and hydrophobic cavities of nuclear hormone receptors (AR, PR, MCR, GCR, SHB, ER, THRA). Well known and characterized drug targets COX2, P38, DHFR and other kinases also allow stable predictions with the method presented here. As expected variability of prediction for druggable proteins is highest for proteins that have average scores around the inflexion point (0.5) of the logistic model. This is for example the case for enoyl reductase, but also phosphodiesterase 5A (and

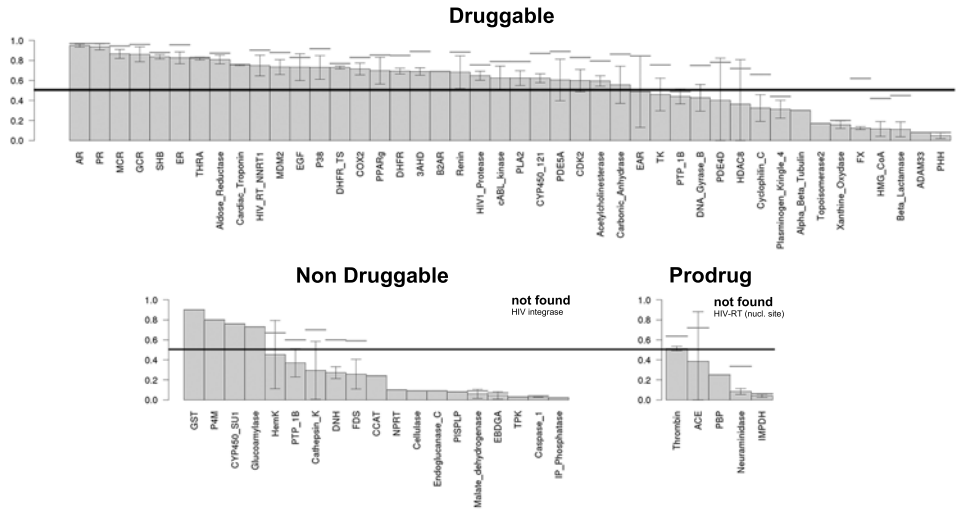


Figure 3.6: Average druggability prediction scores using fpocket on the whole DD. Error bars indicate standard deviations and maximum druggability scores for each target are marked as thick horizontal lines. The theoretical druggability threshold is 0.5.

4D).

A very important prediction variability can be observed for HDAC8. This due to the fact that the cavity detection carried out on a total 10 different cavities yields different results. Thus, for example for HDAC8, maximum druggability scores of 0.72 can be observed, while the average of scores is below the druggability threshold of 0.5. Interestingly on other cavities known to be more solvent exposed and flexible, like P38 and CDK2, more stable predictions are obtained.

AR Androgen Receptor; PR Progesterone Receptor; MCR Mineral Corticoid Receptor; GCR Glucocorticoid Receptor; SHB Sex Hormone Binding globulin; ER Estrogen Receptor; THRA Thyroid Hormone Receptor; DHFR Dihydrofolate Reductase; MDM2 Mouse Double Minute 2-Tumor Protein; EGF Epidermial Growth Factor Receptor; P38 P38 Map Kinase; COX2 Cyclooxygenase 2; DHFR-TS Bifunctional Dihydrofolate Reductase-Thymidylate Synthase; PPARg Peroxisome Proliferator Activated Receptor gamma; B2AR Human Beta 2 Adrenergic Receptor; PLA2 Phospholipase A2; CYP450-51 Cytochrome P450 51; 3AHD 3-Alpha-Hydroxysteroid Dehydrogenase type 3; PDE5A Phosphodiesterase 5; EAR Enoyl Reductase; CDK2 Cell Division Protein Kinase 2; TK Tyrosine Kinase; PTP-1B Protein Tyrosine Phosphatase 1B; PDE4D Phosphodiesterase 4D; HDAC8 Histone Deacetylase 8; FX Factor Xa; HMG-CoA 3-Hydroxy-3-Methylglutaryl-coenzyme A Reductase; PHH P-Hydroxybenzoate Hydrolase; GST Glutathione-S-transferase; P4M Phenylalanine-4-Monooxygenase; CYP450-SU1 Cytochrome P450 105A1; HemK N5-glutamine methyltransferase; DNH Deoxyuridine Nucleotide Hydrolase; FDS Farnesyl Diphosphate Synthase; CCAT cytosolic Branched Chain Aminotransferase; NPRT Nicotinate phosphoribosyltransferase; PISPLP Phosphatidylinositol spec. phospholipase; EBDGA Exo-beta-D-glucosaminidase; TPK Tyrosine protein kinase BTK; ACE Angiotensin Converting Enzyme; PBP Penicillin Binding Protein; IMPDH Inosine Monophosphate Dedhydrogenase

Several cases have been observed where fpocket either detects a complete internal channel system attached to a cavity of interest (ACE) or where no cavity is found (HIV integrase, HIV-RT) with sufficient precision (mutual overlap criterion equals 1). Especially in the context of non druggable cavity prediction, the fact that fpocket does not identify the cavity of interest is in-line with the fact that buried cavities are needed for effective drug binding. The geometric principle exploited by fpocket allows identification of such buried surface patches. If the degree of burial is too low, then fpocket may discard the putative pocket even before scoring.
Other reasons of variabilities during druggability predictions can be pointed out on carbonic anhydrase. The ligand used as reference for indicating if a pocket detected by fpocket is the binding site is given using the residue name. However on several structures the same ligand binds to two different pockets on the same structure. Both cavities yield very different druggability scores. Thus here the prediction seen in the results is altered by the way the data set is used.

Comparison with other methods

A strict comparison of our results with other methods is possible within given limits. Comparing our predictions to the MAP_{POD} score is only feasible on the set of structures used by Cheng and co-workers [Cheng et al., 2007]. Furthermore, our druggability model does not estimate a maximum affinity, but produces a binary measure of druggability. However, as results shown by Cheng can be split up in two different parts, one druggable the other either difficult, or non druggable. Most druggable proteins are classified as such by both, MAP_{POD} and the druggability score. Also for proteins like cathepsin K, caspase 1 (ICE-1), PTP-1B and HIV integrase low scores are predicted or the cavity of interest was deemed too shallow (HIV integrase). For predicting prodrug or difficult binding sites MAP_{POD} clearly allows a ranking of these among non druggable cavities. As the druggability score developed here has a binary behaviour, the average drug score could be situated at around 0.5, but the fact that the inflexion point does not favour scoring close to the threshold. Furthermore, none of the difficult (or prodrug binding protein) pockets have been used for training and validation.
Significant differences between MAP_{POD} can be observed for factor Xa and HMG-CoA reductase. Both are also poorly scored with MAP_{POD}, but considered to be above the druggability threshold. However, in our results factor Xa is predicted non druggable on average. This is mainly due to the fact that the binding site is very heterogeneous and composed of several sub-pockets separated by shallow and solvent exposed areas. Automatic cavity detection mostly allows detection of the S1 sub-pocket of factor Xa. The exact extent of the binding site used for MAP_{POD} calculations for factor Xa is likely to include other sub-pockets. HMG-coA has a very polar binding site allowing drugs like rosuvastatin to make ionic interactions, usually a hallmark for prodrugs.
Another more thorough comparison of our druggability score has been carried out versus Dscore a measure of druggability developed by Schrödinger LLC

using the SiteMap [Halgren, 2007] cavity detection algorithm.

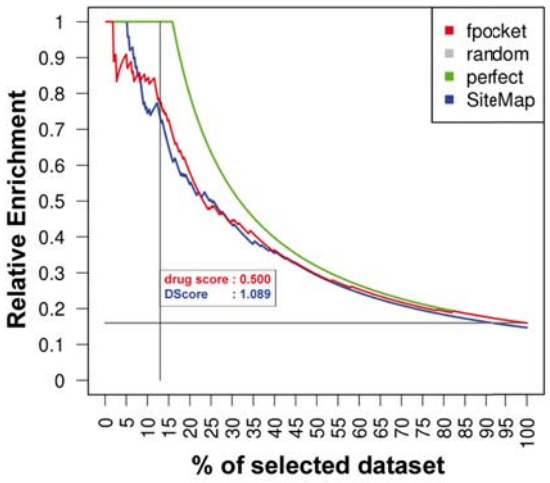


Figure 3.7: Relative enrichment from screening for druggable cavities from a pool of pockets using fpocket and SiteMap

In this study we used the NRDS previously defined as reference data-set. Using automatic cavity detection 430 cavities were identified with SiteMap (63 druggable cavities) versus 440 with fpocket (70 druggable cavities). Not all proteins could be prepared for SiteMap analysis, as the automatic protein preparation protocol failed on 8 structures.
The druggability and Dscore were calculated for all cavities and the resulting list was ordered by each score. A normalized enrichment factor (ratio of druggable cavities in a subset of the set of cavities analyzed) was used to compare the scoring performance of both methods. Results are shown on figure 3.7 where the relative enrichment for each method (fpocket in red, SiteMap in blue) are compared to a perfect prediction case (green) and random prediction (grey). It is noteworthy that both methods yield significantly enriched predictions compared to random ranking of cavities. Very good early enrichment is achieved by SiteMap using Dscore. However, here the binary classification between druggable and non druggable is critical and the thresholds for both scores are depicted on the figure (0.5 for fpocket and 1.089 for SiteMap) as well as their location on the enrichment plot. Both thresholds are situated in the same zone of the plot and show that both methods yield very similar enrichment on this data-set.
These results further confirm that fpocket is as able as SiteMap to identify binding sites and successfully rank them either druggable or non druggable. While producing comparable results, fpocket has various inherent advantages over SiteMap. Instead of searching for cavities in an order of minutes, fpocket performs cavity detection and scoring in a matter of seconds. Fpocket is fully automatized allowing large scale applications without lengthy protein preparation protocols.

Uncertainties on druggability prediction

Results previously summarised on figure 3.6 also showed several proteins, where the cavity was predicted druggable although the status of the protein was assigned as non druggable. This was especially the case for GST, P4M, Glucoamylase and CYP450 Subunit 1. All of these targets were identified in DrugBank as binding a drug. However upon visual inspection they were considered non druggable. P4M was wrongly assigned as target for levodopa in DrugBank. Glutathione-S-transferase and Glucoamylase binders are not drug-like, thus the targets were deemed non druggable. CYP450 and GST are both detoxifying enzymes and therefore binding of drug-like molecules can happen in a non-specific manner, questioning the druggability status of both proteins. While more uncertainties might persist on classifying targets as non druggable, the presence of an actual orally bio-available drug is evidence of the druggability of a target. However, even among known druggable targets, predictions can falsely predict them as non-druggabile or difficult to target. However, upon a more thorough analysis of these targets one can notice that among them, ionic and covalent interactions are driving forces for drug-protein interactions, which is not a hallmark for druggable proteins. This is for instance the case for PHH, ADAM33, β-lactamase or xanthine oxydase. PTP-1B, an interesting system already discussed in a previous work of our group [Seco et al., 2009] yields predictions in the zone of difficult to target pockets.
While assignments of druggability are fixed within this study on the newly established data-set, the Druggable Cavity Directory provides a necessary environment for the scientific community to discuss and alter further assignments and decide on a consensus used throughout the field.

Novelty

In conclusion three principle novelties should be highlighted for this work. First a novel and more comprehensive data-set has been derived to train and evaluate new druggability prediction methods. For the first time systematically different structures for proteins were used. Furthermore apo-structures were considered. This data-set was made available via an interactive online platform.
Second, it is shown for the firs time that polar atoms follow a particular pattern in druggable binding sites. Even if the accessible surface area of polar atoms is low, these were found to make themselves available. This phenomenon is further investigated in the following study.
Third, the work presented here is the first and till today only method to freely and easily detect cavities and estimate their druggability. Last, given the fact that the underlying method is very fast, fpocket with the druggability score can be used as high throughput prediction tool.

Limitations

Main limitations of the method developed here are linked to *(i)* the concept of druggability and *(ii)* the variability of the automated pocket prediction. Druggability is a debatable concept as is the data used to train the druggability

score. For this reason I proposed the druggable cavity directory (DCD), to allow for further discussion and evolution of the current data-set.
An important limitation origins from structural variability. The definition of an automatically identified pocket by fpocket can vary substantially in few cases. As the pocket characteristics used for druggability prediction are derived from such a pocket definition, the outcome is affected.
Last, the prediction is based on training with a specific data-set. Thus a bias of the predictions towards pockets similar to the data-set can be theoretically expected. To accurately assess druggability, low-throughput strategies like MD-mix [Seco et al., 2009] could be considered. However, the druggability score coupled to fpocket is purpose-built to sieve through large structural data-bases and retrieve several systems of which the druggability has to be validated via other techniques (computational and experimental).

Bibliography

Jianghong An, Maxim Totrov, and Ruben Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & cellular proteomics : MCP*, 4(6):752–61, June 2005.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

EE Bolton, Yanli Wang, Paul A Thiessen, and Stephen H Bryant. Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry*, 4(08):217–241, 2008.

Alan C Cheng, Ryan G Coleman, Kathleen T Smyth, Qing Cao, Patricia Soulard, Daniel R Caffrey, Anna C Salzberg, and Enoch S Huang. Structure-based maximal affinity model predicts small-molecule druggability. *Nature biotechnology*, 25(1): 71–5, January 2007.

Eric B Fauman, Brajesh K Rai, and Enoch S Huang. Structure-based druggability assessment-identifying suitable targets for small molecule therapeutics. *Current opinion in chemical biology*, 15(4):468–463, June 2011.

Jianmin Gao, Daryl A Bosco, Evan T Powers, and Jeffery W Kelly. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nature structural & molecular biology*, 16(7):684–90, July 2009.

Philip J Hajduk, Jeffrey R Huth, and Stephen W Fesik. Druggability indices for protein targets derived from NMR-based screening data. *Journal of medicinal chemistry*, 48(7):2518–25, April 2005.

Thomas a Halgren. Identifying and characterizing binding sites and assessing druggability. *Journal of chemical information and modeling*, 49(2):377–89, February 2009.

Tom Halgren. New method for fast and accurate binding-site identification and analysis. *Chemical biology & drug design*, 69(2):146–8, February 2007.

M Hendlich, F Rippmann, and G Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of molecular graphics & modelling*, 15(6):359–63, 389, December 1997.

Tina Hesman Saey. More Than A Chicken, Fewer Than A Grape - Science News, 2010.

R D Hoffmann. Pharmacophores and Pharmacophore Searches Edited by and Rémy D . Hoffmann. *Metabolism Clinical And Experimental*, 32, 2006.

Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature reviews. Drug discovery*, 1(9):727–30, September 2002.

G J Kleywegt and T A Jones. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta crystallographica. Section D, Biological crystallography*, 50(Pt 2):178–85, March 1994.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10:168, January 2009.

O D Monera, T J Sereda, N E Zhou, C M Kay, and R S Hodges. Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. *Journal of peptide science : an official publication of the European Peptide Society*, 1(5):319–29, 1995.

John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature reviews. Drug discovery*, 5(12):993–6, December 2006.

Stéphanie Pérot, Olivier Sperandio, Maria A Miteva, Anne-Claude Camproux, and Bruno O Villoutreix. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug discovery today*, 15(15-16):656–67, August 2010.

Peter Schmidtke, Catherine Souaille, Frédéric Estienne, Nicolas Baurin, and Romano T Kroemer. Large-scale comparison of four binding site detection algorithms. *Journal of chemical information and modeling*, 50(12):2191–200, December 2010.

Jesus Seco, F Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *Journal of medicinal chemistry*, 52(8):2363–71, April 2009.

Michal Vieth, Miles G Siegel, Richard E Higgs, Ian A Watson, Daniel H Robertson, Kenneth A Savin, Gregory L Durst, and Philip A Hipskind. Characteristic physical properties and structural fragments of marketed oral drugs. *Journal of medicinal chemistry*, 47(1):224–32, January 2004.

David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34 (Database issue):D668–72, January 2006.

Entendiendo y Prediciendo la Drugabilidad. Un método High-Throughput para la detección de sitios de unión de fármacos.

Peter Schmidtke, Xavier Barril
Journal of Medicinal Chemistry, 2010, 53 (15)

Las predicciones de drugabilidad son importantes para descartar los targets inabordables y así centrar los esfuerzos en aquellos sitios que, de entrada, ofrecen mejores perspectivas. Sin embargo algunas herramientas de predicción de drugabilidad han sido hechas públicas y ninguna ha sido probada extensamente. Hemos compilado un grupo de cavidades drugables y no drugables en una plataforma colaborativa (`http://fpocket.sourceforge.net/dcd`) que puede ser utilizada, ampliada y corregida con contribuciones comunitarias. Los sitios de unión de fármacos, en ocasiones son sobresimplificados como cavidades cerradas e hidrofóbicas, sin embargo los análisis de datos posteriores revelan que los grupos polares en los sitios de unión de fármacos tienen propiedades que les permiten jugar un papel decisivo en el reconocimiento del ligando. Los datos recogidos, se utilizaron en conjunto con el código abierto fpocket para probar y validar el modelo logístico. El rendimiento punta de esta nueva herramienta ha sido predecir las cavidades de unión de fármacos en sitios de unión conocidos a priori y experimentos de screening virtual, donde se eligieron las cavidades con capacidad de unir fármaco del conjunto inicial. Dicho algoritmo es gratuito, extremadamente rápido y puede ser utilizado de manera efectiva con las cuantiosas colecciones de estructuras.

5858 *J. Med. Chem.* **2010,** *53,* 5858–5867
DOI: 10.1021/jm100574m

Journal of
Medicinal
Chemistry
Article

Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites

Peter Schmidtke*[,‡] and Xavier Barril*[,†,‡]

[†]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Spain, and* [‡]*Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Av. Joan XXIII s/n, 08028 Barcelona, Spain*

Received May 11, 2010

Druggability predictions are important to avoid intractable targets and to focus drug discovery efforts on sites offering better prospects. However, few druggability prediction tools have been released and none has been extensively tested. Here, a set of druggable and nondruggable cavities has been compiled in a collaborative platform (http://fpocket.sourceforge.net/dcd) that can be used, contributed, and curated by the community. Druggable binding sites are often oversimplified as closed, hydrophobic cavities, but data set analysis reveals that polar groups in druggable binding sites have properties that enable them to play a decisive role in ligand recognition. Finally, the data set has been used in conjunction with the open source fpocket suite to train and validate a logistic model. State of the art performance was achieved for predicting druggability on known binding sites and on virtual screening experiments where druggable pockets are retrieved from a pool of decoys. The algorithm is free, extremely fast, and can effectively be used to automatically sieve through massive collections of structures (http://fpocket.sourceforge.net).

Introduction

Despite advances in both experimental and computational fields, it is estimated that around 60% of drug discovery projects fail because the target is found to be not "druggable".[1] Drug discovery project failures are very expensive, and understanding the difficulties associated with a prospective target is essential to balance investment risks. Since the publication of "the druggable genome"[2] and its estimation of the number of therapeutically useful proteins in the human genome, druggability has gradually become part of the target validation process. Traditional target validation tries to assess whether or not alteration of the normal activity of a potential target can have some significant therapeutic effect. The druggability concept adds a structural dimension and evaluates the likelihood that small drug-like molecules can bind a given target with sufficient potency to alter its activity. Several structure-based druggability prediction methods have been published (reviewed in ref 3). Notable contributions in this domain were first done by the group of Hajduk et al.,[4,5] who used NMR-based fragment screening hit rates as a measure of druggability. The model, based on a simple regression analysis, used descriptors like the surface area, the polar/apolar contact area, the roughness, and the number of charged residues in the binding pocket. In 2007, Cheng et al.[6] published a very simple model to estimate the maximum affinity that an ideal drug-like molecule could have for a given binding site. This model was remarkable because correct predictions were obtained on the assumption that binding affinity of drug-like molecules may derive exclusively from the hydrophobic effect. Drug-target molecular recognition is, nevertheless, a more complex phenomenon,[7] and as our understanding of druggability gradually improves, it will influence our view of the drug binding event, just like pharmacokinetics and drug-likeness have influenced each other.[8,9]

With few exceptions,[10] druggability predictions are based on empirical structure−activity relationships, which require a substantial data set on which to train and validate the model. In that regard, the pioneering studies carried out at Abbott and Pfizer are particularly important because they provided an initial pool of test cases which has facilitated subsequent developments.[11] However, these are still limited and, because the druggabilty concept allows for different interpretations, the classification given to some of the targets is debatable. In this study we unify the previous sets and extend them with further examples, adhering to Cheng's et al. definition of the term "druggable", i.e., capable of binding oral drugs. Undoubtedly, the enormous range of binding affinities and bioavailability rates exhibited by this drug class, as well as the fact that some of these molecules are in fact pro-drugs, introduces a fair amount of ambiguity to the definition. Nevertheless, druggability scores will be extremely useful even if they can only provide a qualitative classification between "druggable", "borderline", and "non-druggable". To facilitate further studies, to promote community involvement in the generation of a larger data set, and to reach a wider consensus on the druggability classification, we have made our test set publicly available and editable (http://fpocket.sourceforge.net/dcd).

Initial studies set the path for druggability predictions, but the resulting algorithms were not made available.[4,6] More recently, Schrödinger have used Cheng's data set to endow their SiteMap cavity detection program with a druggability

*To whom correspondence should be addressed. Phone: +34-934031304 (P.S. or X.B.). Fax: +34-934035987 (P.S. or X.B.). E-mail: pschmidtke@ub.edu (P.S.) or xbarril@ub.edu (X.B.).

 Published on Web 07/12/2010

score (Dscore),[11] but it was trained and presented for usage in specific target validation. Thus, the user has to select a very precise zone for performing the druggability prediction. Because the method was trained to estimate druggability of well-defined cavities, it might not be suitable to assess the multitude of cavities that occur in protein structures, many of which have no reported functional role.[12] Automatic predictions on large structural databases would offer the opportunity to identify druggable cavities on sites or targets that might not be considered a priori, such as allosteric sites or proteins for which a therapeutic rationale has not been fully developed. With this goal in mind, the work presented in this article describes a new structure-based target druggability prediction score coupled to the open source cavity prediction algorithm fpocket.[13] We demonstrate the ability of the method to accurately evaluate the druggability of automatically detected cavities, which allows us to correctly rank binding sites both across and within protein structures. At 2 to 4 s per averaged-sized structure, the method is at least 1 order of magnitude more efficient than SiteMap, and the whole of the Protein Data Bank (PDB[a]) can be processed in a few days on a normal computer. fpocket, including the herein presented druggability score, is freely available for download at http://www.sourceforge.net/projects/fpocket.

Last, we discuss which cavity descriptors are more useful to detect druggable cavities and what they tell us about molecular recognition between proteins and drugs. As shown in previous studies, hydrophobicity correlates particularly well with drug binding sites but the implication of a hydrophobicity-explains-all model such as the MAP_{POD}[6] is that the ideal binding site is a closed and "greasy" cavity. Even in the cavities that more closely resemble this description (e.g., the hormone binding site of nuclear receptors), polar interactions play a fundamental role in binding, selectivity, or mediating the biological response.[14] New evidence is presented here for the special characteristics of the polar groups present in binding sites, which provides a more complex and realistic picture of drug−protein molecular association.

Results and Discussion

Compilation of an Open Data Set. A set of protein−oral drug complexes was obtained crossing the list of marketed oral drugs provided by Vieth et al.[15] with the PDB.[16] The DrugBank[17,18] target information was then used to ensure that the complex corresponds to the actual drug-target pair. Visual inspection ensued to classify the complex as druggable, difficult (e.g., in the case of prodrugs), or undruggable (e.g., in the case of nondruglike ligands). Cheng's data set[6] and the public part of Hajduk's data set[4] were added to obtain the druggability data set (DD) used in this study. As most targets are represented by several structures, and to avoid too much bias toward certain protein families, a nonredundant druggability data set (NRDD) was established using a 70% identity cutoff. The composition of the data sets is summarized in Tables 1 and 2.

The notion of druggability is often riddled with uncertainty and classification can be difficult and may evolve over time. A good example of that is the different classification given by Cheng et al. to the related serine proteinases thrombin and factor Xa as "difficult" and "druggable", respectively.[6] As discussed by Halgren, they share common characteristics and it is likely that, eventually, thrombin will become "druggable" although it is an objectively difficult target.[11] For this reason, we deemed necessary to leave the classification of targets open for discussion. Furthermore, we make the data set public in an attempt to instigate participation of scientists from the field into the creation and design of a unified data set. As demonstrated in the docking arena, the establishment of general benchmarks is important to ensure a fair evaluation of prediction performance.[19,20] In the case of druggability predictions, some targets are more easily predicted than others and, even for the same protein, classification may depend on the particular conformation adopted by the receptor. The use of a common data set will facilitate further developments in the field and avoid biases when comparing methodologies. The Druggable Cavity

Table 1. Composition of the DD Data Set; For Comparison, The Composition of the Cheng and Hajduk Data Sets Are Also Shown

	no. of protein structures					
	NRDD[a]	DD[b]			Cheng[c]	Hajduk[d]
druggability classification	holo	holo	apo	total	holo	holo
druggable	45	773	146	919	17 (43)	35
nondruggable	20	75	9	84	4 (10)	37
prodrug	5	60	7	67	6 (10)	
total	70	908	162	1070	27 (63)	72

[a] Nonredundant data set, one structure per protein. [b] Total number of structures in the druggability data set. [c] Reference 6 number of proteins and, in brackets, number of structures. [d] Reference 4 it contains only one structure per protein.

Table 2. Distribution of Structures in DD per Function of the Protein, As Annotated in Uniprot

druggability	no. of structures	function	average resolution
druggable	170	lyase	1.76
	150	hydrolase	1.97
	149	kinase	2.02
	138	nuclear hormone receptor	2.06
	123	oxidoreductase	1.76
	55	Hy[a], Pr[b]	1.98
	42	Hy[a], Pr[b], Tr[c]	2.49
	41	Tr[c], Ox[f]	2.08
	6	structural protein	2.85
	6	transport protein	1.84
	4	ligase	2.17
	4	transferase	2.30
	3	isomerase	1.78
	3	GPCR	3.20
nondruggable	42	Hy[a], Pr[b]	2.12
	21	hydrolase	1.95
	11	transferase	2.32
	4	Lyase	2.18
	3	Tr[c], K_i[d]	2.03
	3	oxidoreductase	1.98
	1	Hy[a], Pr[b], Tr[c], Po[e]	2.10
prodrug binding	39	Hy[a], Pr[b]	2.05
	22	hydrolase	1.85
	3	oxidoreductase	2.82
	3	penicillin binding TMP	2.73

[a] Hydrolase. [b] Protease. [c] Transferase. [d] Kinase. [e] Polymerase. [f] Oxidoreductase.

[a] Abbreviations: ASA, accessible surface area; DD, druggability data set; DCD, druggable cavity directory; MOc, mutual overlap criterion; NRDD, nonredundant druggability data set; PDB, Protein Data Bank; ROC, receiver operator characteristics.

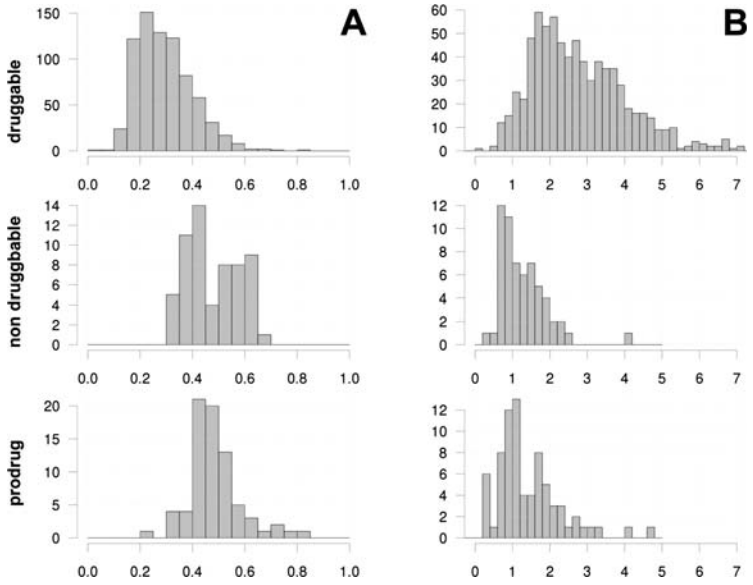


Figure 1. Distribution of (A) the fraction of polar ASA of the pocket and (B) the ASA profile slope ratio on the drugability data set (DD).

Directory (DCD) is a web-based platform that allows multiple users to upload known cavities (identification by PDB code and ligand Hetero Atom Identifier) and assign a value to their druggability in an arbitrary scale from 1 (not druggable) to 10 (druggable). These uploads are then validated by a set of experts (validators) in the field in order to be part of the final data set. Only registered users can upload/validate data. However, anonymous users can access and download the validated data set. The PDB structural database is growing steadily, thus we highly encourage all contributors of the field (medicinal chemists, structural biologists, molecular modellers, etc.) to register, participate, and use this platform, allowing more robust future training/validation cycles for upcoming methods or for retraining of existing methods and scores. The project is available at http://fpocket.sourceforge.net/dcd.

On the Role of Polar Atoms in Druggable Binding Sites. Models with predictive capacity may inform about the physicochemical basis of the underlying process. A good example at hand is the famous rule of five, which predicts drug-likeness based on descriptors related to pharmacokinetics (and very particularly to passive membrane permeation).[21] Similarly, it would be desirable that construction of druggability models could help identify the basis of molecular recognition of drugs by their targets. The Cheng model tells us that the essential feature of a drug binding site is that it should be closed and lipophilic. This was justified on the basis that electrostatic interaction and desolvation energies act in opposition, and the combination of the two is expected to make an insubstantial contribution in the case of charged or polar groups.[6] Nevertheless, the contribution of polar interactions is context dependent, and a single hydrogen bond can contribute as much as 1.8 kcal/mol,[22] comparable to the hydrophobic gain provided by the side-chain of a Val residue.[23] The same applies to ionic interactions.[24] The fact that polar groups play a fundamental role in binding affinity is also supported by the observation that they often constitute anchoring points, featuring predominantly in pharmacophoric models of binding sites.[25] Furthermore, potency is just one of the many factors required for a drug-target complex to result in biological activity. Drugs also have to recognize their target with certain specificity and maintain a stable and specific 3D arrangement within the binding site to be effective. These properties are typically associated with polar interactions, and it seems reasonable to expect that polar groups in binding sites should have some differential properties with regard to the rest of the protein surface. The fact that polar groups are considered irrelevant in Cheng's model or have a negative contribution to druggability in the case of Hajduk et al.[4,6] may wrongly lead to the notion that the ideal druggable site is a completely hydrophobic cavity. We have searched for descriptors with predictive capacity, with the aim of getting an insight on the underlying principles of drug recognition by their targets.

In agreement with previous reports, a predominantly lipophilic composition of druggable binding sites is confirmed in this bigger data set. They typically contain only 20−40% of polar surface versus 40−60% for nondruggable cavities (Figure 1A). Focusing on polar atoms, we find that, on average, 70% of them have very small solvent exposed areas (< 10 Å^2), whereas in nondruggable cavities the proportion decreases to 50%. Considering together the small solvent exposed area of polar atoms and the preponderance of nonpolar atoms, it becomes evident that, in druggable binding sites, protein−ligand hydrogen bonds are surrounded by a hydrophobic environment. In such low dielectric medium, electrostatic interactions become stronger. Very recently, this effect has been quantified in proteins, demonstrating that hydrogen bonds can be up to 1.2 kcal/mol stronger in hydrophobic environments.[26] This clearly indicates that, beyond the obvious gain in hydrophobic potential, a decrease in the polar surface ratio can have the paradoxical effect of increasing the hydrogen bonding potential of the binding site. Without diminishing the importance of hydrophobic interactions, this

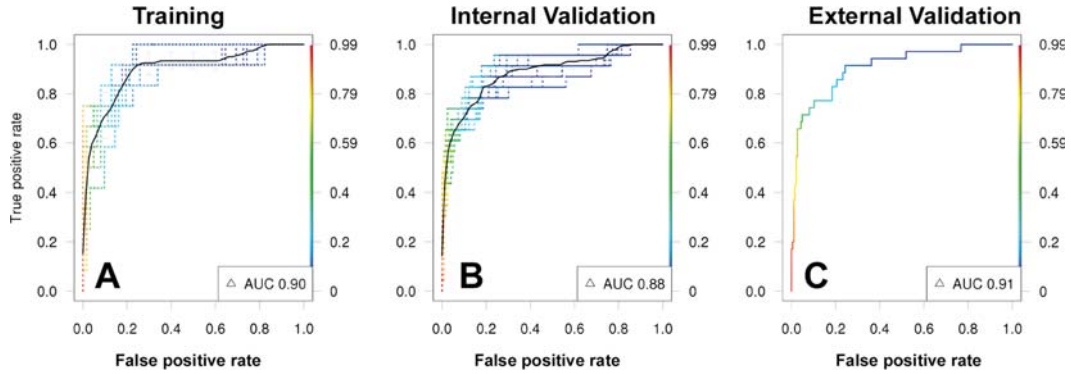


Figure 2. ROC curves for the model building. (A) Training on a total of 74 cavities, (B) internal validation on a total of 146 cavities, (C) external validation on a total of 220 cavities. AUC: area under the curve.

view marries better with the common perception of hydrogen bonds as key elements in drug−protein binding.

Among other descriptors, we investigated the change in accessible surface area (ASA) as a function of the radii used to represent the atoms (see Materials and Methods). Being located in concave regions (cavities), the surface area of a binding site decreases as longer atomic radii are used. Figure 1B shows the ratio of ASA change between polar and nonpolar areas. The different behavior of druggable and nondruggable cavities is highly significant, suggesting that a fundamental aspect of molecular recognition must be associated with this observation. Intriguingly, in nondruggable cavities, the decrease is similar for polar and nonpolar surface areas (average ratio is 1), whereas in druggable cavities, the polar surface area decreases at a much slower rate (average ratio is 3). This means that, in druggable cavities, polar atoms tend to protrude from the cavity surface, making themselves available for interactions (see Figure 6 for a graphical representation). We postulate that the increased protrusion of polar atoms in druggable cavities is fundamental to increase their visibility, rendering them available for interactions and extending the range on which they can exert their selective action. These results also prompted us to investigate the effect of the local environment on the energetics of association of polar groups. In a separate paper (Schmidtke et al., in preparation), we show that the type of local environment found in druggable cavities protects the hydrogen bonds formed between the ligand and the receptor, effectively locking the ligand and permitting longer residence times. Kinetic stability (both of the complex and the binding mode) is another fundamental property of protein−drug complexes that cannot be explained simply on the basis of shape and lipophilic interactions.

Druggability Score. A druggability score was trained and validated on the NRDS, following the protocol described in Materials and Methods. The result of a 10-fold bootstrap run on the learning and internal validation sets is depicted in parts A and B of Figure 2, respectively. The mean prediction result is represented as solid black line, which shows good and stable enrichment. Each cross-validation result is represented as score-colored dashed line. Half of the NDRS was set aside as external validation set, on which the performance of the average model resulting from the learning process was tested (Figure 2C). The resulting scoring function is a two-step logistic function represented by eqs 1−3 and parameters in Table 3 (Materials and Methods). It should be noted that the final formula reflects the need to provide robust predictions in spite of the variability introduced by the automated pocket detection algorithm. Although individually informative, the descriptors in Figure 1 did not yield very robust models during the 10-fold bootstrap learning and validation procedure. These ASA-based descriptors gather information on atomic detail, but as the fpocket cavity prediction can be rather variable from one structure to another or from one conformer to another, their corresponding values are not sufficiently consistent. Instead, the most important descriptor in terms of predictive performance is the mean local hydrophobic density of the binding site. This descriptor combines size and spatial distribution of hydrophobic subpockets into a single number. The two other descriptors used to predict druggability are the hydrophobicity and normalized polarity scores, both of which refer to the physicochemical character of the amino acids lining the pocket. It is noteworthy that both contribute favorably to the score, indicating that both hydrophobic and polar residues can make the binding site more druggable. Again, the residue-based character of theses descriptors makes them more granular, but they also show far less variability than atom-based descriptors during the construction of the final scoring function. As one of the main aims of this scoring function is its high-throughput applicability, it must deliver robust predictions in spite of the varying definition of the binding site provided by the automatic pocket detection algorithm. Thus we would like to emphasize that the scoring function reflects the application for which it has been designed. Descriptors such as those described in Figure 1 are more interpretable and provide a finer level of detail but may be better suited for supervised applications such as predicting the druggability of known binding sites.

As the score is indented for high-throughput and fully automatic predictions, it is necessary to assess its robustness across different crystal structures of the same target. With that purpose, the scoring function established with the NRDS was applied to the whole of the DD. This is a major difference with previous methods. Hajduk et al.[4] used a single structure per protein, whereas Cheng et al.[6] used a variable number of structures, but with no apparent logic. For instance, p38 MAP kinase, for which multitude of structures are available in the PDB, is represented by a single structure (1KV1; DFG-out conformation). Figure 3 illustrates the average score and the standard deviation for each target in the data set. Satisfactorily,

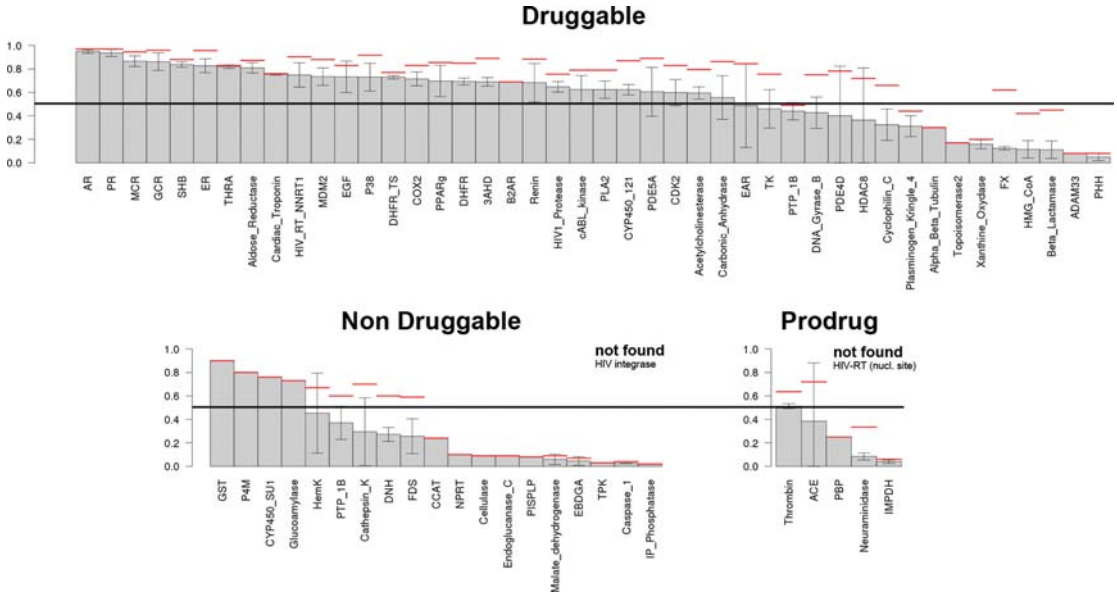


Figure 3. Prediction of druggability on all structures of the DD. Error bars correspond to mean prediction ± standard deviation. For details and full list of protein name abbreviations, refer to Supporting Information.

the standard deviation is low at both ends of the distribution. Variability is larger for scores approximating 0.5, but this is a natural consequence of using a logistic model. Another cause for variability is the pocket prediction itself. In the case of buried and fairly rigid cavities such as the nuclear hormone receptors, the pocket detection algorithm produces consistent results and the druggability score is very stable. In solvent exposed or flexible cavities, the automated pocket detection may yield significantly different binding site predictions and the druggability score may diverge. HDAC8 or CDK2 exemplify this situation, as manifested by very large standard deviations and individual predictions ranging from nondruggable to very druggable. For instance, the average drug score for HDAC8 is 0.36, but the best scoring cavity gets a value of 0.72, comfortably within the druggable range. On rare occasions, the pocket detection algorithm may completely fail to identify the drug binding site. On ACE, fpocket detects the whole internal channel system as one single and continuous cavity, thus it is too large (up to 7700A^3) to consider it as a proper definition of the binding site. For very shallow binding sites, such as the HIV integrase, a cavity may not be detected at all. Arguably, this is not a bad result but a mere reflection of the nondruggable character of the site. In summary, variability in the cavity definition step is an intrinsic limitation of the method presented here, but constraining the druggability prediction to a very concise zone around the experimentally known binding pocket would forbid large scale applicability of the method. Notwithstanding this limitation, the results in Figure 3 and Table S1 (Supporting Information) demonstrate the predictive performance of druggability measurements, indicating that fpocket most often produces consistent results and that the drug score formula manages well pocket variability.

The fact that information in the PDB is very often redundant is an advantage for the method, as multiple predictions can be obtained for a given site. Druggability predictions can then be based on the average score (with a cutoff of 0.5), but if the associated standard deviation is large, it may be advisible to take the value for the top scorers instead. Inspection of Figure 3 (red lines) indicates that, in this case, a cutoff value of 0.7 provides better discriminating power between the druggable and nondruggable sets.

Protein flexibility may also be a source of variability. For instance, the PDE4D binding site can be exposed to solvent or it may be closed due to interactions between the UCR2 domain and the catalytic domain. Interestingly, this conformational change may influence druggability, as a recent paper[27] shows that allosteric modulators binding to the former state have improved side effects compared to known PDE4D inhibitors. All binding sites containing the allosteric modulator were scored (six pockets originally not included in the test set), yielding a higher and less variable drug score than the open conformations (0.55 ± 0.12 and 0.4 ± 0.28, respectively). Another system, not originally included in the druggability data set, for which conformational changes have been associated to different degrees of druggability is renin.[28] We have therefore calculated the druggability on structure 2BKS, which contains two monomers. Chain A has no inhibitor and adopts the conformation that, as stated by Davis et al., is less suitable for drug discovery.[28] Nevertheless, it should be noted that aliskiren is an approved drug targeting this conformation.[29] On chain B, ligand binding induces opening of a hydrophobic subpocket (Figure 4), allegedly making it more druggable. The score obtained was 0.9 and 0.93, respectively. So in this case the structural rearrangements do not alter significantly the druggability prediction, which suggests that the difficulty to develop drugs from the closed conformation may be related to the chemical scaffold of the inhibitors (peptidomimetics) rather than to the physical−chemical properties of the binding site.

Another protein yielding variable predictions is the druggable carbonic anhydrase. However, here the prediction error is due to the fact that many structures contain twice the same ligand, one in the actual binding site and a second

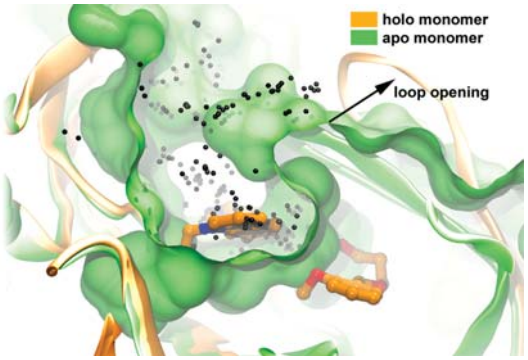


Figure 4. Renin druggable binding site yielding similar druggability scores for holo and apo monomers.

one in a more superficial cavity (e.g., PDB codes 2QOA, 2NNS). As the ligand is used to identify the binding pocket, two completely different binding sites are thus scored. Satisfactorily, they yield opposing drug scores. Accordingly, the variability seen on carbonic anhydrase is not a limitation of the method but of the data set used to evaluate the method.

Comparison with MAP_{POD}. Analyzing the results obtained for the structures in the Cheng data set, few noticeable differences can be observed with the MAP_{POD} score.[6] Proteins like HIV RT (NNRTI site), COX2, CDK2, MDM2, CYP450_121, EGF, PDE5A, acetylcholinesterase, p38, and others are clearly classified as druggable by both methods. Nevertheless, the scores of one protein relative to another do not correlate between models, reflecting the fact that our score is a binary classifier whereas the MAP_{POD} value aims at predicting the maximal binding affinity for the binding site. Agreement is also obtained for the nondruggable proteins cathepsin K, caspase 1 (ICE-1) and PTP-1B, which yield rather low druggability scores. As mentioned above, HIV Integrase has a very shallow binding site that is not even identified by fpocket, indicating that the binding site is not buried enough to wrap a small molecule.

Regarding proteins binding prodrugs, results show a clear separation between druggable binding sites and such "difficult" to target binding sites, again in consonance with the MAP_{POD} score. Considering that druggability score was trained as a bimodal predictive model and none of the prodrug binding sites was used during training or validation, classification of proteins in this category in the nondruggable class is a desirable behavior of the model. Particularly encouraging is the case of thrombin, which receives a score of 0.5. As discussed by Halgren,[11] dabigatran etexilate, a prodrug targeting thrombin, is now marketed in Europe and Canada. Approval by the U.S. Food and Drug Administration is pending for this year.

There is an apparent discrepancy between druggability score and MAP_{POD} on proteins Factor Xa and HMG-CoA reductase. But, in fact, both proteins also yield rather low MAP_{POD} scores, 100-fold lower than the following protein in the druggable data set (DNA gyrase B). The low scores obtained by both methods on HMG-CoA reductase can be explained by the very polar nature of the binding site, which forms ionic interactions with the drugs (e.g., rosuvastatin). Factor Xa, on the other hand, has a heterogeneous and partly shallow cavity, which could influence the low MAP_{POD} score. In the present study, the low score results from the binding site identification protocol, which yields two distinct pockets instead of one.

In conclusion, the method presented here is able to reproduce results obtained by previous publications while not focusing on the binding site of interest, which is a crucial requirement for automated high-throughput druggability predictions.

Wrong Predictions or Wrong Druggability Status? Sometimes, target misclassification can be directly attributed to the ambiguous nature of the "druggability" concept. In the case of druggable cavities with very low score, the pockets are usually very small or host ionic interaction patterns (P-hydroxybenzoate hydrolase, ADAM33). In other cases, the protein forms covalent bonds with the ligand (β lactamase and xanthine oxidase). In those cases, druggability is largely the result of a specific chemical feature rather than a global property of the cavity. Correct assignment of these cavities may therefore require a completely different approach to the one used here. For large-scale predictions, these failures should not be significant, as the sites correspond to enzymatic catalytic centers whose relevance is already evident and can be detected by other means.[30] A borderline druggability score for the allosteric binding site of PTP-1B can hardly be considered a failure given that these inhibitors are weak[31] and the binding site offers limited opportunities for tight binding.[10]

Good druggability scores for nondruggable targets GST, P4M, cytochrome P450 105A1, and glucoamylase, can be traced to the fact that all of these proteins do have well-defined binding sites. They were initially selected because they are listed in the DrugBank as targets of an approved drug and the drug−target complex is available in the PDB. Upon detailed analysis, it was decided that the targets could not really be considered as druggable and were included in the nondruggable data set with the aim of improving the balance between positive and negative data. Reasons to reverse the original classification included wrong drug-target assignment in the DrugBank (e.g., P4M was listed as a target of levodopa), lack of drug-likeness on the ligand part (e.g., the GST inhibitor ethacrynic acid), or the biological role (e.g., GST and CYP are detoxifying enzymes). The druggability status of these proteins is therefore debatable. Glucoamylase is known to bind the oral drug acarbose with high affinity,[32,33] but because this drug acts in the intestine and is not bioavailable, the original classification as nondruggable by Hajduk et al.[34] seems adequate. In all these cases, SiteMap also predicts the sites as druggable or difficult, never as nondruggable. The DCD provides an adequate environment to reconsider the classification of these and other proteins in the data set.

Druggability Prediction on Apo Structures. Because of induced-fit effects, holo structures may present a different arrangement of the features that determine binding. In the molecular docking field, holo structures have been shown to provide qualitatively better results.[35] As training and druggability predictions have been carried out on holo structures, it was necessary to test the robustness of predictions on apo structures to ensure that unused druggable cavities can be found when screening large structural databases (Figure 4). As shown on Table S1 (Supporting Information), the druggability score observed in holo structures is generally reproduced within the range of confidence in apo structures. The main source of variability appears to be the number of apo

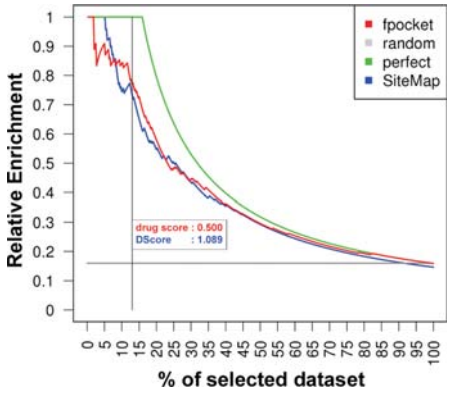


Figure 5. Comparison of druggability prediction performance between fpocket druggability score and SiteMap Dscore.

structures, which is comparatively small. In fact, there are only two proteins with a large number of apo structures: carbonic anhydrase and β lactamase. The latter is predicted as nondruggable, and the average value is identical for the apo and holo sets (0.11). The former is interesting because it is the most populated set (83 holo and 61 apo structures), and the results suggests that there is some information decay as the apo structures get worst average values than holo (0.4 ± 0.3 and 0.5 ± 0.3 respectively). Nevertheless, the best apo cavities get druggability scores as high as the best holo structures (close to 0.9). In the rest of the systems, the number of apo structures is insufficient to get statistics, but the individual values fall within the range observed with holo structures. β-2-Aadrenergic receptor/T4 lysozyme chimera is an exception to this rule, as the apo structures (PDB codes: 2R4R, 2R4S) have missing residues on the extracellular side of the transmembrane helices, where the catecholamines bind. This demonstrates that, except in the case of grossly different cavity shapes, the predictions are sufficiently robust to detect apo as well as holo binding sites.

Comparison with SiteMap. Finally, the method presented here was compared with Dscore, a recent contribution by Thomas Halgren from Schrödinger,[11] which is, to the best of our knowledge, the only other software available to screen for druggable cavities out of the box. The performance of both scores was assessed on the druggable and nondruggable structures from the NRDD. The druggability score is intimately linked to the procedure used to define the cavity; therefore cavities are not interchangeable between programs and have to be generated independently. SiteMap analysis was run on 63 structures, as the remaining eight gave problems when running SiteMap in an automated way. Fixing these structures would have required manual intervention, but this was not done because the goal was to simulate an automated large-scale screen for druggable cavities. As a result, the number of cavities differs between methods (70 druggable, 440 total for fpocket; 63 druggable, 430 total for SiteMap). The performance metric used to compare the methods must take into account the different composition of the data set. We have used a normalized enrichment factor, defined as the ratio of druggable cavities in a given subset of the library. Figure 5 plots this value versus the amount of selected cavities after ordering them by decreasing druggability score/DScore. In both cases, the enrichment factor is ideal in the beginning and remains very good throughout. Predictive power of the druggability score decreases sharply after 0.5, which is the expected behavior, as this value corresponds to the inflection point in the logistic model. The corresponding Dscore at the same fraction of the library is 1.1, coinciding with the average value for druggable cavities in the Dscore training set.[11]

In terms of performance, both methods are similarly capable of retrieving druggable cavities from structural databases. Nevertheless, fpocket present two important advantages for large scale screening purposes. First, the method is very fast (1−2 s for structures up to 450 residues compared to a few minutes for SiteMap). Second, fpocket is completely automatic and does not need protein preparation or selection of parameters. Additionally, the logistic scoring scheme provides a natural cutoff for acceptable enrichment values.

Conclusions

Considerations about druggability are becoming part of the target selection process (see, for instance, ref 36). If the structure is available, this can be done by visual inspection of the binding site but, in the absence of clear guidelines, the decision may be largely subjective. Compiling a large set of targets with their associated druggability is an efficient way of making sure that previous knowledge is retained, thus contributing to our understanding of the fundamental processes behind druggability. Here we present the largest druggability data set to date, which the community can freely download, edit, or extend. In comparing druggable to nondruggable binding sites we find that, contrary to previous models,[6] hydrophobicity is not the sole determinant of target druggability, as polar groups also play an important role on the recognition of drug-like molecules. The data set can also be used to train computational methods or to assess their performance. Previously, computational methods were trained to predict the druggability of known binding sites. Here we have placed particular emphasis on the ability of the program to automatically detect binding sites and subsequently assess their druggability. The resulting software is state of the art in terms of druggability prediction performance while having the advantage of being free of charge, open source, and computationally very efficient. As the druggability prediction is directly associated to a cavity detection method, screening for druggable cavities in large structural data sets is straightforward. Application of this method to the PDB could, then, provide insights into the druggable targetome already contained in the structural proteome. Putative drug binding sites can then be further analyzed by complementary methods.[3,10,37] We expect that this will be a useful approach to unlock promising yet largely unpursued mechanisms of action such as allosteric modulation,[38] protein−protein inhibitors,[39] pharmacological chaperones,[40] or interfacial inhibitors.[41]

Materials and Methods

Data Set. Currently, one data set for assessment of druggability prediction methods was commonly used. This data set was provided by Cheng et al.[6] and was further used for validation of SiteMap druggability score.[11] For a more robust validation of this method, a bigger and nonredundant data set is provided. The herein presented data set was derived using a study published in 2004 by Vieth et al.[15] In this paper, characteristic physical properties and structural fragments of marketed oral

drugs were derived from an extensive data set. Only orally available marketed drugs were kept from this data set.

Next, PubChem (pubchem.ncbi.nlm.nih.gov) was used to check whether a 3D structure of the drug in whatever protein exists in the PDB.[16] Only drugs having resolved 3D structures were kept. Next, data was crossed with DrugBank[17,18] entries for these drugs. The DrugBank contains entries of the targets corresponding to each drug. The known 3D structures from PubChem Compound, corresponding to the actual target of the drug, were kept for further analysis. Finally, structures were checked by hand to establish whether the drug in the protein could be classified as drug, prodrug, or if the binding site should be considered undruggable due to missing drug likeness of the ligand. Structures from Cheng's data set were also added, and the same set was enhanced by other structures for the same proteins, resulting in the druggability data set (DD). Generally, crystal structures with a resolution lower than 2.5 Å and R_{free} below 0.3 were kept despite some exceptions for underrepresented protein classes.

These steps allowed building up of an extensive data set containing for major parts druggable proteins. However, known negative information is also very important and very difficult to find in this field. To enhance the data set with known negative data, the nondruggable proteins from Cheng's data set and parts of the data set published by Hajduk and co-workers[4] was used. Table S1 in Supporting Information summarizes the contents of the data set as well as the prediction results.

Last, for druggable, prodrug binding and nondruggable proteins, a nonredundant data set (NRDD) was established using the BlastClust clusters from the PDB at a maximum of 70% sequence similarity (ftp://ftp.wwpdb.org/pub/pdb/derived_data/NR/clusters70.txt). For learning and validation, only well-defined known druggable and nondruggable cavities were chosen. This was done using the fpocket mutual overlap criterion (MOc),[13] that allows assessment if a found cavity covers well the actual ligand binding site or not. Thus, known druggable and nondruggable cavities needed a MOc of 1 in order to be considered for learning and validation of a druggability scoring function.

Cavity Detection. For this study, fpocket, a highly scalable and free open source pocket detection software package, was used.[13] Extensive usage was made especially of the dpocket program, allowing easy extraction of pocket descriptors. The default set of dpocket descriptors was extended by polar and apolar pocket surface area (van der Waals surface +1.4 Å and van der Waals surface +2.2 Å). The dpocket derived pocket descriptors were further tested for suitability in the creation of a druggability score using logistic regression.

Finally, the retained pocket descriptors are:

- The normalized mean local hydrophobic density. This descriptor tries to identify if the binding pocket contains local parts that are rather hydrophobic. For each apolar α sphere the number of apolar α sphere neighbors is detected by seeking for overlapping apolar α spheres. The sum of all apolar α sphere neighbors is divided by the total number of apolar α spheres in the pocket. Last, this score is normalized compared to other binding pockets on the same protein.
- The hydrophobicity score. This descriptor is based on a residue based hydrophobicity scale published by Monera et al.[42] For all residues implicated in the binding site, the mean hydrophobicity score is calculated and is used as descriptor for the whole pocket. Each residue is evaluated only once.
- The normalized polarity score. As published on http://www.info.univ-angers.fr/~gh/Idas/proprietes.htm, each residue can be split in two polarity categories (1 and 2) The final polarity score is the mean of all polarity scores of all residues in the binding pocket. Each residue is evaluated only once.

Table 3. Constants of Druggability Score Model

descriptor	coefficient	mean value[a]	standard deviation/mean[b]
intercept	ss_0	−6.238	−0.095
mean local hydrophobic density (normalized)	ss_1	4.592	0.154
hydrophobicity score	ss_2	5.717	0.170
polarity score (normalized)	ss_3	3.985	0.459
intercept	$ss_{1,0}$	−5.141	−0.170
mean local hydrophobic density (normalized)	$ss_{1,1}$	6.579	0.173
intercept	$ss_{2,0}$	−2.669	−0.168
hydrophobicity score	$ss_{2,1}$	0.056	0.216
intercept	$ss_{3,0}$	−2.445	−0.238
polarity score (normalized)	$ss_{3,1}$	2.762	0.330

[a] Mean values refer to mean constants after a 10-fold bootstrap run. [b] The ratio between the standard deviation and the mean value of the constant assess the variability of the constant during the bootstrap.

Druggability Score. As the NRDD contains only 45 druggable and 21 nondruggable proteins, the following rule was considered: all other cavities (not in contact with a ligand) identified on the proteins containing at least one druggable cavity and having a size higher than 60 α spheres (corresponding to reasonably sized cavities) were also considered as nondruggable. This rule allows to introduce decoys into the given set of cavities in the NRDD and thus increase its size substantially.

To train and validate the druggability score, the NRDD, consisting of 70 druggable cavities, 16 nondruggable cavities having a MOc of 1, and 354 decoys was split in two. The first half of the data was used to train the model. Training and internal validation was performed using a 10-fold bootstrap with a one-third/two-third training/validation ratio. First logistic models were derived for each pocket descriptor using the glm function from the R statistical software package,[43] and according to predictive power and stability during the 10-fold bootstrap, these models were further considered.

In the next step, predictions coming from these "one descriptor based" logistic models were associated in one common logistic model where statistically nonsignificant descriptors or unstable models were filtered out. The general model is shown in eqs 1−3 as drugscore, the single descriptor based models are designated by the function $f_x(d_x)$, where d_x is a given descriptor.

$$\text{drugscore}(z) = \frac{e^{-z}}{1+e^{-z}} \tag{1}$$

$$z = \beta_0 + \beta_1 f_1(d_1) + \beta_2 f_2(d_2) + \beta_3 f_3(d_3) \tag{2}$$

$$f_x(\mathrm{d}_x) = \frac{e^{-\beta_{x,0}+\beta_{x,1}d_x}}{1+e^{-\beta_{x,0}+\beta_{x,1}d_x}} \tag{3}$$

The coefficients of the model, as shown in Table 3, were obtained by averaging coefficients derived on each step of the bootstrap.

The second half of NRDD was reserved for external validation (selected by random before the bootstrap run). Receiver operator characteristics (ROC) and derivative figures shown in this paper were produced using the ROCR package.[44]

Comparison with Schrödinger SiteMap. To compare fpocket druggability prediction performance with SiteMap Dscore, the NRDD data set was used as benchmark. SiteMap was systematically launched on all structures after running the Prepwizard protein preparation protocol of Maestro. SiteMap was run to accept a maximum number of 10 binding sites to reproduce prediction results published by Halgren. However, no binding site restrictions were applied. A binding site was successfully recognized if at least 20 site points were less than 1.5 Å away

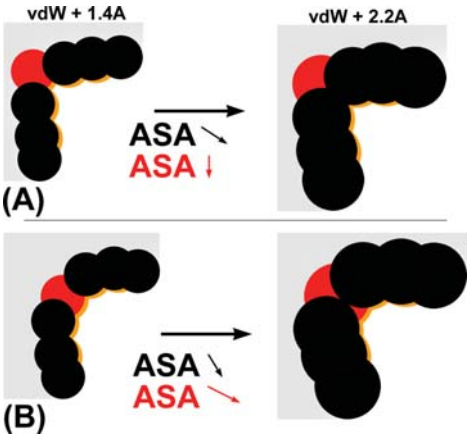


Figure 6. Principles of ASA calculation upon atom fattening. Apolar atoms are colored in black, polar in red, the ASA is shown in orange: (A) a buried polar atom has a low ASA when taking into account the van der Waals radius +1.4 Å. This ASA disappears if one increases the probe size to 2.2 Å. (B) A more exposed polar atom has also a low ASA when taking a van der Waals radius +1.4 Å but has still a contribution to the total ASA of the pocket when taking the van der Waals raidus +2.2 Å.

from any of the atoms of the known ligand. Unlike the training/validation procedure, all other binding sites on druggable and nondruggable proteins were considered as decoys for both algorithms.

For this comparison, fpocket was run on the NRDD and the druggability score calculated for each binding site. A binding site was considered druggable/nondruggable according the rules specified in the previous part of Materials and Methods.

Pocket Surface Calculations. To assess the importance of polar atoms in known drug binding sites (bound structure), surface calculations were performed in the following way. The set of pocket atoms were identified as the atoms within at most 5.5 Å from the nearest ligand atom. For this set of atoms, the portion of van der Waals surface nonoccluded by surrounding atoms was calculated.

To calculate an ASA close to the solvent-accessible surface area, the van der Waals radius of each pocket atom was increased by 1.4 Å for the surface calculation.

Next, atom fattening was performed, increasing the van der Waals radius correction from 1.4 to 2.2 Å in steps of 0.1 Å. For each modified van der Waals radius, the van der Waals surface of the pocket was calculated. By definition, in concave portions of the protein surface, the van der Waals surface decreases upon atom fattening. In the protein binding sites assessed here, these surfaces decreased linearly. Because of this linear behavior in the radius range considered (van der Waals +1.4 Å to van der Waals +2.2 Å), an automatic construction of linear regression based models is possible and reliable. Thus for the polar and apolar ASA profiles, two models with two parameters (slope and intersect) for each can be obtained. The behavior of these parameters was assessed throughout this study, enabling direct access to the concavity of the pocket for polar and apolar atoms (slope) and the ratio of polar versus apolar atoms (ratio of intersects). The principle is shown in Figure 6. Furthermore, the ratio between the apolar and polar slope of the concavity profiles is of importance in this work and will be referenced to concavity profile ratio. The concavity profile calculation was implemented for running on NVIDIA graphics processing units using python, the excellent Biskit structural bioinformatics framework,[45] and PyCUDA.[46]

Acknowledgment. We thank Vincent Le Guilloux for careful proofreading of the manuscript. This work was financed by the Ministerio de Educación y Ciencia (grant SAF2009-08811). P.S. is funded by the Generalitat de Catalunya. The "Servei de Disseny de Fàrmacs" at CESCA is acknowledged for access to commercial software.

Supporting Information Available: Full list of the results obtained for each protein in the data set. List of protein name abbreviations used in Figure 3. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Brown, D.; Superti-Furga, G. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today* **2003**, *8*, 1067–1077.
(2) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nature Rev. Drug Discovery* **2002**, *1*, 727–730.
(3) Egner, U.; Hillig, R. C. A structural biology view of target drugability. *Expert Opin. Drug Discovery* **2008**, *3*, 391–401.
(4) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **2005**, *48*, 2518–2525.
(5) Hajduk, P. J.; Huth, J. R.; Tse, C. Predicting protein druggability. *Drug Discovery Today* **2005**, *10*, 1675–1682.
(6) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnol.* **2007**, *25*, 71–75.
(7) Böhm, H. J.; Klebe, G. What can we learn from molecular recognition in protein−ligand complexes for the design of new drugs? *Angew. Chem., Int. Ed. Engl.* **2003**, *35*, 2588–2614.
(8) Lajiness, M. S.; Vieth, M.; Erickson, J. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 470–477.
(9) Vistoli, G.; Pedretti, A.; Testa, B. Assessing drug-likeness—what are we missing? *Drug Discovery Today* **2008**, *13*, 285–294.
(10) Seco, J.; Luque, F. J.; Barril, X. Binding site detection and druggability index from first principle. *J. Med. Chem.* **2009**, *52*, 2363–2371.
(11) Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.
(12) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892–906.
(13) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.
(14) Sack, J. S.; Kish, K. F.; Wang, C.; Attar, R. M.; Kiefer, S. E.; An, Y.; Wu, G. Y.; Scheffler, J. E.; Salvati, M. E.; Krystek, S. R., Jr; Weinmann, R.; Einspahr, H. M. Crystallographic structures of the ligand-binding domains of the androgen receptor and its T877A mutant complexed with the natural agonist dihydrotestosterone. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4904–4909.
(15) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **2004**, *47*, 224–232.
(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(17) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.
(18) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–6.
(19) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(20) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.
(21) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
(22) Fersht, A. R. The hydrogen bond in molecular recognition. *Trends Biochem. Sci.* **1987**, *12*, 301–304.

(23) Karplus, P. A. Hydrophobicity regained. *Protein Sci.* **1997**, *6*, 1302–1307.
(24) Barril, X.; Aleman, C.; Orozco, M.; Luque, F. J. Salt bridge interactions: stability of the ionic and neutral complexes in the gas phase, in solution, and in proteins. *Proteins* **1998**, *32*, 67–79.
(25) Langer, T.; Hoffmann, R. D. Pharmacophores and Pharmacophore Searches. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH: Weinheim, Germany, 2006; Vol. 32, pp 375.
(26) Gao, J.; Bosco, D. A.; Powers, E. T.; Kelly, J. W. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nature Struct. Mol. Biol.* **2009**, *16*, 684–690.
(27) Burgin, A. B.; Magnusson, O. T.; Singh, J.; Witte, P.; Staker, B. L.; Bjornsson, J. M.; Thorsteinsdottir, M.; Hrafnsdottir, S.; Hagen, T.; Kiselyov, A. S.; Stewart, L. J.; Gurney, M. E. Design of phosphodiesterase 4D (PDE4D) allosteric modulators for enhancing cognition with improved safety. *Nature Biotechnol.* **2010**, *28*, 63–70.
(28) Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem., Int. Ed. Engl.* **2003**, *42*, 2718–2736.
(29) Staessen, J. A.; Li, Y.; Richart, T. Oral renin inhibitors. *Lancet* **2006**, *368*, 1449–1456.
(30) Sankararaman, S.; Sha, F.; Kirsch, J. F.; Jordan, M. I.; Sjolander, K. Active site prediction using evolutionary and structural information. *Bioinformatics* **2010**, *26*, 617–624.
(31) Wiesmann, C.; Barr, K. J.; Kung, J.; Zhu, J.; Erlanson, D. A.; Shen, W.; Fahr, B. J.; Zhong, M.; Taylor, L.; Randal, M.; McDowell, R. S.; Hansen, S. K. Allosteric inhibition of protein tyrosine phosphatase 1B. *Nature Struct. Mol. Biol.* **2004**, *11*, 730–737.
(32) Sierks, M. R.; Svensson, B. Functional roles of the invariant aspartic acid 55, tyrosine 306, and aspartic acid 309 in glucoamylase from *Aspergillus awamori* studied by mutagenesis. *Biochemistry* **1993**, *32*, 1113–1117.
(33) James, J. A.; Lee, B. H. Glucoamylases: microbial sources, industrial applications and molecular biology: a review. *J. Food Biochem.* **1997**, *21*, 1–52.
(34) Andrews, J. S.; Weimar, T.; Frandsen, T. P.; Svensson, B.; Pinto, B. M. Novel Disaccharides Containing Sulfur in the Ring and Nitrogen in the Interglycosidic Linkage. Conformation of Methyl 5′-Thio-4-*N*-alpha-Maltoside Bound to Glucoamylase and Its Activity as a Competitive Inhibitor. *J. Am. Chem. Soc.* **1995**, *117*, 10799–10804.
(35) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **2003**, *46*, 4265–4272.
(36) Campbell, S. J.; Gaulton, A.; Marshall, J.; Bichko, D.; Martin, S.; Brouwer, C.; Harland, L. Visualizing the drug target landscape. *Drug Discovery Today* **2010**, *15*, 3–15.
(37) Panjkovich, A.; Daura, X. Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC Struct. Biol.* **2010**, *10*, 9.
(38) Lindsley, J. E.; Rutter, J. Whence cometh the allosterome? *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 10533–10535.
(39) Arkin, M. R.; Wells, J. A. Small-molecule inhibitors of protein−protein interactions: progressing towards the dream. *Nature Rev. Drug Discovery* **2004**, *3*, 301–317.
(40) Leandro, P.; Gomes, C. M. Protein misfolding in conformational disorders: rescue of folding defects and chemical chaperoning. *Mini Rev. Med. Chem.* **2008**, *8*, 901–911.
(41) Pommier, Y.; Cherfils, J. Interfacial inhibition of macromolecular interactions: nature's paradigm for drug discovery. *Trends Pharmacol. Sci.* **2005**, *26*, 138–145.
(42) Monera, O. D.; Sereda, T. J.; Zhou, N. E.; Kay, C. M.; Hodges, R. S. Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. *J. Pept. Sci.* **1995**, *1*, 319–329.
(43) R Development Core Team. *R: a language and environment for statistical computing*, **2007**.
(44) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **2005**, *21*, 3940–3941.
(45) Grunberg, R.; Nilges, M.; Leckner, J. Biskit—a software platform for structural bioinformatics. *Bioinformatics* **2007**, *23*, 769–770.
(46) Klöckner, A.; Pinto, N.; Lee, Y.; Catanzaro, B.; Ivanov, P.; Fasih, A. PyCUDA: GPU Run-Time Code Generation for High-Performance Computing. *arXiv* **2009**.

Supporting Information

Understanding and predicting druggability. A high-throughput method for detection of drug binding sites.

Peter Schmidtke, Xavier Barril

Contents:

List of protein name abbreviations used in Figure 3.

AR : Androgen Receptor

PR : Progesterone Receptor

MCR : Mineral Corticoid Receptor

GCR : Glucocorticoid Receptor

SHB : Sex Hormone Binding globulin

ER : Estrogen Receptor

THRA : Thyroid Hormone Receptor

DHFR : Dihydrofolate Reductase

MDM2 : Mouse Double Minute 2-Tumor Protein

EGF : Epidermial Growth Factor Receptor

P38 : P38 Map Kinase

COX2 : Cyclooxygenase 2

DHFR-TS : Bifunctional Dihydrofolate Reductase-Thymidylate Synthase

PPARg : Peroxisome Proliferator Activated Receptor gamma

B2AR : Human Beta2 Adrenergic Receptor

PLA2 : Phospholipase A2

CYP450_51 : Cytochrome P450 51

3AHD : 3Alpha-Hydroxysteroid Dehydrogenase type 3

PDE5A : Phosphodiesterase 5

EAR : Enoyl Reductase

CDK2 : Cell Division Protein Kinase 2

TK : Tyrosine Kinase

PTP-1B : Protein Tyrosine Phosphatase 1B

PDE4D : Phosphodiesterase 4D

HDAC8 : Histone Deacetylase 8

FX : Factor XA

HMG_CoA : 3-Hydroxy-3-Methylglutaryl-coenzyme A Reductase

PHH : P-Hydroxybenzoate Hydrolase

GST : Glutathione S-Transferase

P4M : Phenylalanine-4-Monooxygenase

CYP450_SU1 : Cytochrome P450 105A1

HemK : N5-glutamine methyltransferase

DNH : Deoxyuridine Nucleotide Hydrolase

FDS : Farnesyl Diphosphate Synthase

CCAT : cytosolic Branched Chain Aminotransferase

NPRT : Nicotinate phosphoribosyltransferase

PISPLP : Phosphatidylinositol spec. phospholipase

EBDGA : Exo-beta-D-glucosaminidase

TPK : Tyrosine protein kinase BTK

ACE : Angiotensin Converting Enzyme

PBP : Penicillin Binding Protein

IMPDH : Inosine Monophosphate Dedhydrogenase

Table S1. Summary of prediction results per protein.

	Holo cavities					Apo cavities				
Protein Name	No. of cavities	Mean score	Std. dev	Mean score of top10%	Confi-dence[1]	No. of cavities	Mean score	Std. dev	Mean score of top10%	Confi-dence[1]
druggable										
Carbonic Anhydrase	76	0.56	0.31	0.86	0.93	61	0.42	0.30	0.82	0.83
DHFR-TS	3	0.73	0.06	0.77	0.90	-				
DHFR	39	0.69	0.14	0.85	0.91	1	0.73	0.00	0.73	0.00
Thymidine Kinase	29	0.46	0.19	0.76	0.87	2	0.43	0.06	0.47	0.87
SHB	6	0.84	0.03	0.88	0.96	-				
PDE4D	54	0.40	0.28	0.78	0.84	-				
Mineralcorticoid Receptor	25	0.87	0.09	0.95	0.94	-				
Glucocorticoid Receptor	6	0.86	0.07	0.96	0.91	-				
Androgen Receptor	38	0.95	0.01	0.97	0.99	-				
Progesterone Receptor	9	0.94	0.03	0.97	0.97	-				
Estrogen Receptor	77	0.83	0.10	0.96	0.95	2	0.86	0.01	0.87	0.97
CYP P450 - 121	13	0.62	0.21	0.87	0.82	-				
HIV RT - NNRTI	38	0.75	0.09	0.90	0.94	-				
HIV1 Protease	24	0.65	0.14	0.76	0.90	1	0.44	0.00	0.44	0.00
PPARγ	43	0.70	0.12	0.86	0.93	2	0.50	0.04	0.52	0.93
ADAM33	1	0.08	0.00	0.08	0.00	1	0.15	0.00	0.15	0.00
Acetylcholinesterase	42	0.60	0.26	0.80	0.84	4	0.68	0.11	0.74	0.85
HMG CoA Reductase	27	0.12	0.00	0.42	0.93	-				
DNA gyrase B	3	0.43	0.28	0.75	0.53	-				
Plasminogen Kringle 4	8	0.31	0.15	0.44	0.84	1	0.29	0.00	0.29	0.00
p-hydroxybenzoate Hydroxylase	12	0.05	0.02	0.08	0.98	-				
COX2	9	0.72	0.11	0.83	0.89	-				
Phospholipase A2	19	0.62	0.19	0.79	0.85	4	0.77	0.04	0.80	0.94

Beta-2-adrenergic Receptor/T4 lysozyme chimera	1	0.69	0.00	0.69	0.00	2	0.25	0.21	0.39	0.56
Enoyl-[acyl-carrier-protein] reductase	22	0.49	0.26	0.85	0.81	2	0.47	0.07	0.47	0.85
Coagulation factor X,	20	0.12	0.18	0.62	0.87	-				
Cardiac Troponin C	2	0.76	0.01	0.76	0.99	-				
3 Alpha Hydroxysteroid Dehydrogenase	12	0.69	0.27	0.89	0.76	-				
Beta Lactamase	59	0.11	0.12	0.45	0.93	20	0.11	0.09	0.34	0.93
PDE-5A	17	0.61	0.22	0.89	0.82	2	0.74	0.13	0.83	0.73
THRA protein	2	0.82	0.01	0.83	0.97	-				
Cyclophilin C	4	0.33	0.24	0.66	0.66	-				
Aldose Reductase	67	0.81	0.08	0.87	0.96	5	0.90	0.06	0.86	0.93
cABL Kinase	17	0.63	0.19	0.79	0.85	-				
CDK2	50	0.59	0.27	0.83	0.84	2	0.32	0.20	0.46	0.59
Epidemial Growth Factor	3	0.73	0.10	0.83	0.83					
MDM2	5	0.73	0.18	0.88	0.77					
P38 Map Kinase	43	0.73	0.21	0.92	0.87	2	0.49	0.27	0.91	0.44
Renin	37	0.68	0.20	0.88	0.95	1	0.88	0.00	0.88	0.00
HDAC	10	0.36	0.28	0.72	0.73					
Topoisomerase 2	1	0.17	0.00	0.17	0.00					
Xanthine Oxydase	4	0.10	0.02	0.12	0.97					
Alpha Beta Tubulin	1	0.30	0.00	0.30	0.00					
PTP-1B (druggable site)	2	0.44	0.07	0.49	0.85					
non druggable										
cytosolic Branched Chain Aminotransferase 1	1	0.24	0.00	0.24	0.00	-				

Farnesyl Diphosphate Synthase	3	0.26	0.29	0.59	0.51	-				
Glutathione S-Transferase	1	0.90	0.00	0.90	0.00	-				
Nicotinate phosphoribosyltransferase	1	0.10	0.00	0.10	0.00	-	0.00	0.00	0.00	0.00
Exo-beta-D-glucosaminidase	2	0.05	0.04	0.07	0.93	-	0.00	0.00	0.00	0.00
Phenylalanine-4-Monooxygenase	1	0.80	0.00	0.80	0.00	-	0.00	0.00	0.00	0.00
CYP P450 – SU1	1	0.76	0.00	0.76	0.00	-	0.00	0.00	0.00	0.00
HIV RT Nucl. site	-	0.00	0.00		0.00	-	0.00	0.00	0.00	0.00
Cathepsin K	14	0.30	0.25	0.70	0.79	-	0.00	0.00	0.00	0.00
PTP-1B (non druggable site)	6	0.37	0.26	0.60	0.69	-	0.00	0.00	0.00	0.00
Caspase 1	8	0.03	0.01	0.04	0.99	-	0.00	0.00	0.00	0.00
HIV Integrase	1	0.01	0.00		0.00	-	0.00	0.00	0.00	0.00
HemK	7	0.45	0.22	0.67	0.76	-	0.00	0.00	0.00	0.00
IP Phosphatase	1	0.02	0.00	0.02	0.00	-	0.00	0.00	0.00	0.00
Tyrosine protein kinase BTK	2	0.03	0.00	0.03	0.00	-	0.00	0.00	0.00	0.00
Deoxyuridine nucleotide hydrolase	5	0.27	0.19	0.60	0.76	-	0.00	0.00	0.00	0.00
Glucoamylase	1	0.73	0.00	0.73	0.00	-	0.00	0.00	0.00	0.00
Phosphatidylinositol spec. phospholipase	1	0.08	0.00	0.08	0.00	3	0.05	0.03	0.07	0.96
Endoglucanase C	1	0.09	0.00	0.09	0.00	-	0.00	0.00	0.00	0.00
Malate dehydrogenase	2	0.06	0.04	0.09	0.92	-	0.00	0.00	0.00	0.00
Prodrug binding										
IMPDH	6	0.04	0.01	0.35	0.99	-	0.00	0.00	0.00	0.00
Neuraminidase	22	0.08	0.10	0.34	0.92	3	0.05	0.03	0.07	0.95
Thrombin	30	0.51	0.18	0.64	0.88	-	0.00	0.00	0.00	0.00

Penicillin Binding Protein	1	0.25	0.00	0.25	0.00	-	0.00	0.00	0.00	0.00
Angiotensin converting enzyme	2	0.39	0.47	0.72	0.01	-	0.00	0.00	0.00	0.00

[1]The confidence level is a measure of possibility of accurate prediction. The confidence level is calculated as difference between 1 and the ratio of the standard deviation of the drug score and the log of the number structures incremented by 1. Thus a confidence level close to 1 signifies that drug score yielded very stable results on multiple structures. The less structures one uses or the more the prediction gets variable in one protein structural family, the lower the confidence.

3.2 Structure kinetics relations

3.2.1 Introduction

As described in the previous section on druggability, an intriguing relation was found between druggability and the relative exposure and protrusion of polar atoms in the rather apolar environment of druggable binding sites. The analysis of several druggable versus non druggable binding sites revealed that druggable binding sites tend to contain polar atoms that are less exposed (70% of all polar atoms with a surface area below $10\mathring{A}^2$). Furthermore, the relative protrusion of these poorly exposed atoms into the lumen of the pocket is higher than in non druggable pockets, meaning that these polar atoms despite their low surface area, tend to make themselves more available for an interaction. The study described in this section was undertaken to investigate this relationship of polar atoms with respect to the druggability of a given binding site. This lead us to investigate binding kinetics, an aspect of protein-ligand interactions that is poorly understood, but increasingly perceived as a fundamental aspect controlling the biological activity of drugs.
Thermodynamics reign in drug discovery and binding affinity optimisation, even though kinetic aspects are central and would provide a more detailed vision of the binding process [Holdgate and Ward, 2005, Swinney, 2009, Englert et al., 2010].
David C Swinney, a pioneer in promoting the importance of binding kinetics in drug discovery describes the current situation in the following poignant terms: *"Current capabilities for the use of binding kinetics in shaping and effecting safe, therapeutic responses are rudimentary. There is a lack of understanding or awareness that binding kinetics can be used to differentiate drugs, as well as a lack of appreciation that the optimization of binding kinetics and mechanisms can drive an increase in the therapeutic index of a drug. Furthermore, the application of binding kinetics as a discovery and optimization parameter currently requires an empirical approach due to the challenge of employing this reductionist biochemical tool to complex physiological systems."*
Figure 3.8 illustrates basic concepts of energetics of an equilibrium binding / unbinding process where higher barriers are correlated with slower exchange rates (A). In several studies Swinney shows the potential of considering kinetics in particular in addition to the understanding of underlying mechanisms of action [Swinney and Anthony, 2011] to define drug efficacy, safety, duration of action and differentiation to other medicines. This effort is supported by Copeland [Copeland et al., 2006] and Zhang [Zhang and Monsma, 2009], focusing particularly on the effect of long residence times of drug molecules in their respective binding sites, directly linked to the duration of action and thus pharmacokinetic profiles, target selectivity and drug safety. The importance of residence time was already acknowledged very early in a study by Leysen and Gommeren on serotonin-S_2, dopamin-D_2 and histamin-H_1 antagonists and opiates[Leysen and Gommeren, 1986]. In the context of increasing importance and possibilities in systems biology approaches Ohlson shows the putative interest of transient drugs, characterised by fast off and on rates [Ohlson, 2008]. Despite these contributions, the pharmaceutical industry *"still appears to be acting with*

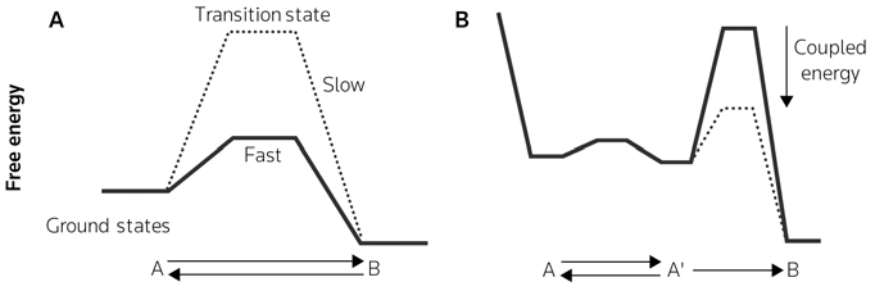


Figure 3.8: Figure and legend adapted from [Swinney, 2009]: **A:** Energetics of a simple equilibrium reaction; the position of the equilibrium is determined by the relative energy between A and B, while the height of the barrier describes the rate of transition between the two states. Equilibrium with a high barrier is reached slowly, whereas equilibrium with a low barrier is achieved rapidly. **B:** Boundaries to accessible states are determined by the height of the energy barrier. Solid line; A and A' are in equilibrium and B is not accessible because of a high energy barrier. Dashed line; the barrier to B is lowered by energy provided by a coupled system such as drug binding, enzyme catalysis and induced conformational changes. This is an example of a kinetically controlled reaction.

trepidation toward assessing the fundamental role of kinetics in drug actions" [Zhang and Monsma, 2009]. Reasons for this hesitating behaviour of industry can be manifold, but a central aspect is the lack of understanding of binding kinetics and the lack of known structural characteristics on the ligand and receptor site influencing kinetics.

Observing drug binding

A very privileged group of researchers around David E Shaw lately started to harvest impressive success of the long term project around very long molecular dynamics simulations. Using a specifically developed computer architecture called ANTON, and MD software Desmond, the group was able to observe spontaneous drug binding in various proteins, like the β_1 and β_2 - adrenergic receptors [Dror et al., 2011] or Src kinase [Shan et al., 2011]. Such simulations can help to understand at molecular level which events are hindering and favouring binding of a molecule on a specific target. Interestingly but not surprisingly, water has been found to play a central role in ligand binding events. Running such extensive molecular dynamics simulations is currently not possible for academics labs, especially outside the US. A cheaper alternative is to use software optimized for GPUs. Buch et al. have used this approach (ACEMD) to produce 495 replicas of 100 ns MD trajectories, observing spontaneous binding of benzamidine to β-trypsin in 187 simulations [Buch et al., 2011]. These pioneering works should prove useful in understanding structure kinetic relationships and awake interest in pharmaceutical industry. However, unbiased MD simulations can currently only be used to study fast processes. Thus, there is a need for alternatives for studying slower processes, such as drug-target dis-

sociation (drugs on the market have residence times varying from seconds to days).

Structure kinetic relationships

Using properties observed during the work on druggability, we intended to derive putative relationships between the structure of proteins and the kinetics of ligand association and dissociation. As both are intimately linked to solvation and desolvation, the role of water molecules appears to be crucial for mediating binding and unbinding processes.

3.2.2 Objectives

Analyse the influence of the pocket properties on ligand binding The initial and main objective guiding this work was to understand the observation that poorly exposed polar atoms seem to protrude more into the pocket lumen in druggable binding sites. In order to do so, the druggability has to be analysed with respect to 3 parameters: *(i)* the role of the surface area of the polar atom, *(ii)* the role of the cavity shape in the local environment of the polar atom and *(iii)* the role of the protrusion of polar atoms into the cavity.
From this main objective various other conclusions and further investigations have been undertaken to prove the findings experimentally.

ABPA and water molecules A subsequent objective was related to the investigation of the role of almost buried polar atoms (ABPAs) to the stability or slow kinetics of water molecules.

ABPAs and drug design A last objective of this study was to prove that the structure-kinetics relationship observed on waters and small ligands translates to protein-ligand binding and that the principle can be used to understand and predict structure kinetic relationships (SKRs).

3.2.3 Results

The structural analysis of real biological systems like proteins is very complex. The surface shape and distribution of different atoms is rather heterogeneous. Thus, analysing the effect of the pocket curvature and the exposure of one polar atom in an apolar environment on a protein is not trivial.
To simplify the problem and still be able to seek the role of protruding polar atoms in apolar environments, we created an artificial cavity, which intends to mimic the main properties we are trying to investigate within this study. This artificial cavity is built up by a truncated restrained methane half sphere. A cyanonitril molecule acting as hydrogen bond acceptor is positioned at the bottom of this hydrophobic concave shape. Last, the system is solvated with TIP3P water and ammonia is placed in the lumen of the artificial pocket acting as hydrogen bond donor and ligand.

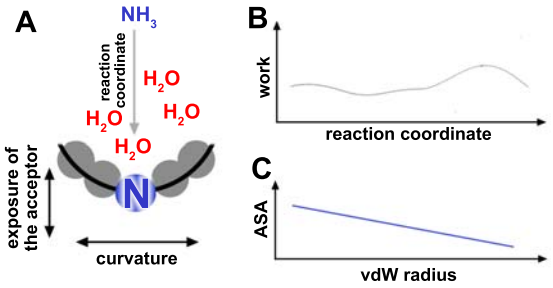


Figure 3.9: Overview of the theoretical experiment. **A:** Architecture of the artificial cavity of methane (grey spheres) and cyanonitrile as H-bond acceptor (N). Ammonia is approached (reaction coordinate) to the H-bond acceptor. **B:** While approaching ammonia to the cavity, the necessary work is tracked. **C:** For each combination of curvature and exposure of the acceptor atom the accessible surface area is calculated using increasing van der Waals radii for each atom of the system.

The advantage of such an artificial system is that the curvature of the cavity as well as the protrusion of the hydrogen bond acceptor can be freely modulated to assess very different scenarios otherwise difficult to analyse. Using the Jarzynski relation detailed in Materials & Methods equation 2.16 we can relate the work necessary to approach ammonia towards the hydrogen bond acceptor to the free energy. The overall system architecture is schematically shown on figure 3.9
Steered molecular dynamics (SMD) have been run on this system for various configurations of cavity curvature and polar atom exposure (29 in total). During 20 replicas for each configuration the ammonia has been slowly (4.5$\mathring{A}$ per ns) approached to the hydrogen bond acceptor in the bottom of the cavity. We found that, when the polar atom had specific structural characteristics a

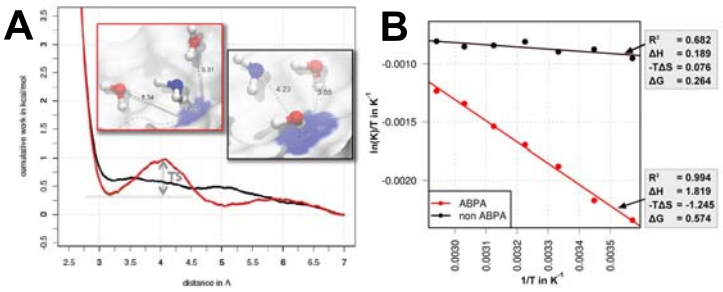


Figure 3.10: **A:** Free energy profiles of ammonia approaching an hydrogen bond acceptor, (red) when this acceptor is an ABPA and (black) when the acceptor is not an ABPA. When the hydrogen bond acceptor is an ABPA a clear transition state can be observed (TS). **B:** Van't Hoff analysis of an ABPA and a non-ABPA system.

transition state as shown on figure 3.10 A (red curve) can be observed, when the ligand is situated between the first and the second solvation shell. More

precisely, the transition state exists only when :

1. $2\mathring{A}^2 < A_0 < 10\mathring{A}^2$
2. when $\Delta A < 0$

where A_0 is the accessible surface area of the H-bond acceptor in the bottom of the cavity and ΔA the difference of accessible surface area of the H-bond acceptor using fattened and default van der Waals radii for the surface calculation of the cavity (see figure 3.9 C). The polar atoms following these characteristics will be designated as almost buried polar atoms, or *ABPA's*. To our knowledge, this is the first time that such simple structural characteristics can be linked to a kinetic property like the one observed here.
The presence of the transition state corresponds to a situation during the approach of ammonia when the hydrogen bond acceptor in the bottom of the cavity has to be desolvated. As represented in figure 3.10 A, this desolvation happens in a concerted-like manner when the polar atom is well exposed and very accessible. However, for ABPA's the acceptor has to be previously desolvated to finally allow ammonia to approach the acceptor to an ideal hydrogen bond distance. This intermediate state explains the energetic penalty measured by the transition state.
To assess if the appearance of the transition state is enthalpic or entropic, two configurations of the cavity were chosen (one containing the polar atom as an ABPA, another one with a more exposed polar atom) and SMD's were run for seven different temperatures producing 100 replicas for each system and temperature. These were subsequently used to derive the entropic and enthalpic terms of the free energy doing a van't Hoff analysis. Figure 3.10 B shows that the intermediate lengthening of the hydrogen bond between the ABPA and the ligand results in an enthalpic penalty, partially offset by a positive entropic contribution.
These results show that we can link structural properties of the cavity to a kinetic effect due to a specific desolvation pattern upon binding and unbinding of the ligand. Even if the artificial system that was used here to study these properties is very simplistic compared to complex biological systems, it allowed us to derive the elementary role of an almost buried atom and showed its link to kinetic properties of ligand binding. As previously shown on figure 3.10, the binding of a ligand can be energetically neutral. However, the presence of an ABPA can act as kinetic trap lengthening the residence time of the ligand within the binding site.

Dispersion of water molecules

While these theoretical studies allow us to demonstrate what might happen on an elementary level, evidence on proteins has to be found to support this hypothesis. ABPA's can be found all over the protein surface. However local shape and hydrophobicity might play a key role in the relative importance of one ABPA versus another on the protein surface. Given the wealth of experimental structures available in the PDB, the stability of water molecules on the

protein surface was further analysed. Applying the kinetic trapping or ABPA theory, water molecules hydrogen bonded to ABPA's on the protein should be more stable.
During this analysis different measures of stability have been used to evaluate if this hypothesis could be verified. First the radial distribution of water around backbone carbonyles of the protein has been analysed on a set of 2.704.956 atom pairs. On figure 3.11 A the radial distribution of water molecules is

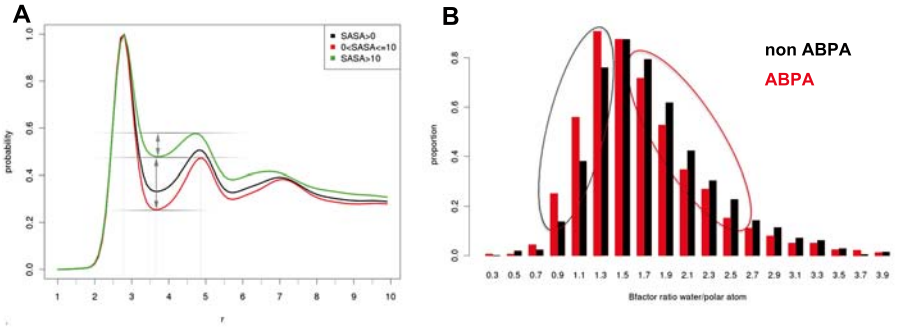
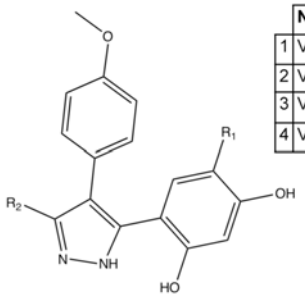


Figure 3.11: **(A)** Radial distribution of waters around backbone carbonyles for different accessible surface areas of the carbonyles. **(B)** Ratio of B-factors between waters and carbonyles on the protein.

shown around the backbone carbonyles of proteins. The black curve corresponds to all carbonyles considered, the red curve only to carbonyles having an accessible surface area in agreement with the definition of an ABPA. The green curve shows atoms having larger accessible surface areas (¿10$\mathring{A}$). The radial distribution is normalised by the maximum probability of finding water in the first solvation shell (first peak at around 2.6$\mathring{A}$). However, the interesting result is the well-depth between the first and the second solvation shell between ABPA's and exposed atoms. Clearly, solvation shells are better defined for ABPA's than for other atoms which could indicate that water molecules are more stable around ABPA's. To foster data from crystallographic structures, electron density maps were gathered for 1723 different proteins from the Uppsala Electron Density Server [Kleywegt et al., 2004]. Next, the electron density of water molecules making a single hydrogen bond and being in an apolar environment was analysed. Figure 3.11 B shows the B-factors of these water molecules normalized by the B-factors of the carbonyle of the protein. Here again water molecules in contact with ABPA's tend to be less mobile.
Both studies suggest that ABPA's can act as kinetic traps, as suggested by our theoretical model. Further validation of the theory has been undertaken by analysing a set of unrelated protein MD trajectories. Water stability around ABPA's was investigated and it was found that the long residence times of water are exclusive of ABPA's. Thus these particular polar atoms are a necessary but not sufficient condition for stabilising water on the protein surface.

ABPA's in protein ligand complexes

Like previously shown on figure 3.10, exchange of water molecules is energetically neutral. Thus, the presence of a transition state like the one observed when interacting with an ABPA is directly linked to longer residence times of waters with ABPA's. However, if one considers molecules that are usually object of drug discovery the situation is far more complex. While the kinetic effect described here might still take place, it can be difficult to extricate from other events happening upon association and dissociation of the compound. Still, the concept presented in this work was investigated on a *real life* example of a series of 4 compounds, shown on figure 3.12 of known activity on HSP90. These compounds are from the resorcinol series which is currently in Phase I and II of clinical trials. All chosen compounds are very similar. This is mainly to minimise possible problems explained previously.



	Name	R1	R2
1	VER37655	-Br	-Me
2	VER49181	-Br	-CONHEt
3	VER49008	-Cl	-CONHMe
4	VER49009	-Cl	-CONHEt

Figure 3.12: Overview of the 4 compounds analysed in this study

Interacting with an ABPA More precisely, these 4 compounds are composed of a set of 2 pairs of compounds that have been chosen as extensive crystallographic data is available on HSP90. The first pair of compounds (compound 1 & 2) will be used as a test case for benefits that one can expect when interacting with an ABPA on the protein. While compound 2 is able to hydrogen bond to the backbone carbonyle of G97 via an added amide moiety in R2, compound 1 is unable to do so. This oxygen on G97 is an ABPA, having an accessible surface area of around $4\mathring{A}^2$ and a $\Delta A < 0$. The differences between

	ITC(25°C)			SPR(25°C)			SPR(279K-303K)		
Compound	ΔG	ΔH	$-T\Delta S$	K_{on}	K_{off}	ΔG	ΔG^*_{off}	ΔH^*_{off}	$-T\Delta S^*_{off}$
1	-9.01	-1.54	-7.47	4.3e+05	2.2e-02	-9.94	19.8	21.4	-1.7
2	-10.32	-2.74	-7.58	4.6e+06	6.1e-03	-12.09	20.4	15.6	4.9
Diff / *(ratio)*	1.3	1.2	0.1	(0.09)	(3.5)	2.1	-0.7	5.8	-6.6

Table 3.3: ITC and SPR data for compound 1 & 2. Differences & ratios are significantly different (5% risk) expect the entropy derived via ITC.

compound 1 and 2 are quite substantial, structurally and with respect to experimental results. Analysis of the later reveals that compound 2 interacts more favourably with HSP90 than compound 1 ($\Delta\Delta G$=1.3 kcal/mol). Interestingly, the half life is also 3.5 times longer for the compound making a hydrogen bond with the ABPA.
On the theoretical model established using artificial cavities a particular thermodynamic signature was found: $\Delta H > 0$, $\Delta S > 0 \rightarrow \Delta G > 0$. In the more complex example here, the enthalpy of the transition is however nearly 6 kcal/mol larger for compound 1 and entropy is negative for compound 2.
In order to analyse more deeply the reasons for this disagreement between experimental results and the theoretical model, here again steered molecular dynamics simulations have been performed. This time however, the ligand was pulled out of the binding site (initiation of dissociation). Figure 3.13 is

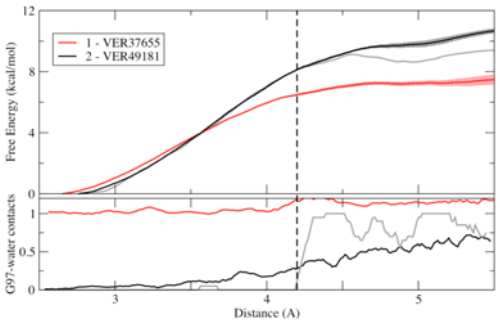


Figure 3.13: Free energy profiles (top) of compounds 1 and 2 derived from SMD simulations. The bottom shows the number of water contacts with the ABPA on G97 during the dissociation pathway.

composed of two parts. The top graph shows the free energy versus the distance between the center of mass of the compound and a reference point in the binding site. The bottom plot shows average number of water contacts of the ABPA while in contact with either compound 1 or 2. On both plots it is clearly visible *(i)* that compound 1 (red) dissociates faster (lower slope of energy while pulling the ligand out) and that the ABPA is always in contact with 1 water molecule. On the contrary, when interacting with compound 2, the ABPA is not solvated in the beginning of the simulation, but gradually gets solvated during the dissociation process. The grey curve is the result of one SMD simulation of compound 2 with a free energy profile close to average. Here, one can clearly see the solvation starting (one water molecule is interacting) when the slope of the energy profile changes at around 4$\mathring{A}$ of dissociation pathway. The change of slope in the energy profiles indicate hydrogen bond breaking events and it can be seen that these events do not occur at the same moment for both compounds (earlier for compound 1 than for compound 2). This emphasizes the stabilizing role the interaction with the ABPA has. Furthermore it gives insights into molecular details of the dissociation process. Indeed it was found that both compounds dissociate differently, explaining the discrepancies between experimental results and the theoretical model. The hydrogen bond

added via compound 2 is found to be very stabilizing playing an important role in the stability of the compound. During dissociation, the resorcinol ring initiates dissociation even before the hydrogen bond made with G97 is broken. On the contrary in compound 2 such a stabilization is not present, thus the methyl group can move unhindered out of the pocket.

Shielding and ABPA The first example investigated the benefits one could expect targeting specifically ABPAs during a possible lead-optimisation process. However it also showed how difficult it is to extricate the kinetic contributions to the overall behaviour of the compound. Using a second set of compounds, another aspect is studied while interacting with ABPAs on the protein surface. Here, both compounds interact with the carbonyle of G97. The only difference is that compound 3 has a methyl attached to the amide, while compound 4 contains an ethyl. Figure 3.14 shows both compounds in

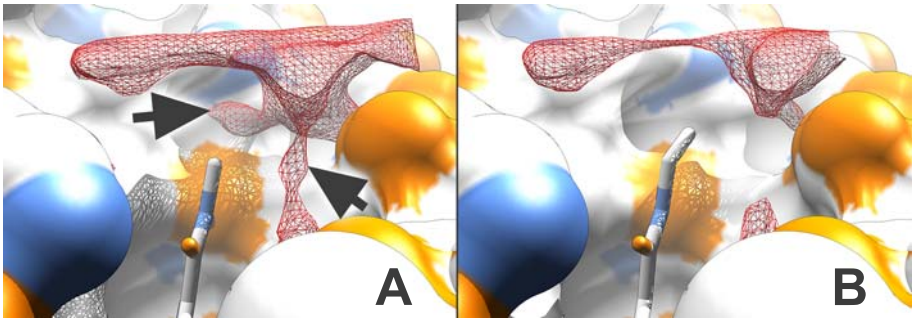


Figure 3.14: Detail on the binding site and compounds 3 (A) and 4 (B) interacting with the ABPA. Water densities at 75% occupancy are shown. Arrows indicate significant changes in hydration pattern in (A).

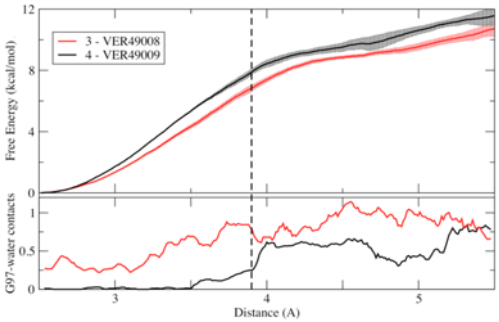


Figure 3.15: Free energy profiles (top) of compounds 3 and 4 derived from SMD simulations. The bottom shows the average number of water contacts with the ABPA on G97 during the dissociation pathway.

the binding site and their respective modifications. The water density observed during a 50ns MD trajectory is represented as red mesh at 75% occupancy. Interestingly, here the added ethyl group acts as protector of the hydrogen bond between the ABPA and the ligand. This can be clearly observed on part A of

the figure (indicated by arrows), where water can be seen on favourable positions to attack the protein-ligand complex.
Steered molecular dynamics were then run for compounds 3 and 4, pulling the ligands out of the binding site. Figure 3.15 shows the free energy profile with respect to the position along the dissociation path for compound 3 (red) and compound 4 (black). Here subtle but visible differences can be noted with respect to the slope of the energy profile. Compound 4 dissociates less easily than compound 3. In the lower part of the plot, data already observed on figure 3.14 is confirmed. Here we can see that the ABPA is on average more solvated along the MD trajectory when it is interacting with compound 3. Experimental results (table 3.4) obtained via ITC and SPR show that the

	ITC(25°C)			SPR(25°C)			SPR(279K-303K)		
Compound	ΔG	ΔH	$-T\Delta S$	K_{on}	K_{off}	ΔG	ΔG^*_{off}	ΔH^*_{off}	$-T\Delta S^*_{off}$
3	-10.13	-3.79	-6.35	1.4e+06	1.12-02	-11.13	20.3	20.5	-0.2
4	-10.44	-3.69	-6.74	9.9e+05	6.9e-03	-11.10	20.6	21.1	-0.5
Diff / *(ratio)*	0.3^b	-0.1^c	0.4^c	$(1.4)^{c,d}$	$(1.6)^{b,d}$	0.0^c	-0.3^a	-0.6^b	0.3^c

Table 3.4: ITC and SPR data for compound 3 & 4. Significant difference with p-value: $^a \leq 0.01$, $^b \leq 0.05$; No significant difference: $^c \geq 0.01$; ^{d} significance calculated from ratio values over temperature range

methyl to ethyl modification has little effect on the binding affinity, improving it by 0.3 kcal/mol when adding the ethyl group (only measurable via ITC). More noticeable is the difference in rate constants. Both compounds have the same k_{on} but compound 4 has a 1.6 times slower k_{off}. Even more striking is the fact that the thermodynamic signature derived from theoretical systems is observed within this example. Altogether, SMD simulations and experimental results suggest that shielding the hydrogen bond made with the ABPA is beneficial for increasing residence time of the compound. Another striking result is the fact that shielding appears to primarily influence dissociation and not the association of the compound. The dissociation process starts with disruption of hydrogen bonds, linked to the solvation of these. On the contrary, during the association process the site can already be pre-desolvated via other interactions with the ligand.

In the context of druggability The initial aim of this study was to relate the observation of particular polar atoms, ABPAs, to the fact that cavities are druggable or not. In the work presented here, the relation between ABPAs and a specific kinetic behaviour has been observed. Considering the kinetic properties that ABPAs might have, especially in an apolar environment, and given that druggable cavities are more apolar than non-druggable cavities it is fair to assume that kinetics do play a role in the tractability of a cavity with a non covalent neutral drug-like molecule. Interestingly, non-druggable cavities, either make covalent interactions with the ligand or they tend to contain charged residues, allowing to make ionic interactions. On protein-protein complexes, such interactions are known to alter association rates but not k_{off}. The

influence of ABPAs observed here however, is likely to lengthen the residence time of the molecule, relating longer residence times to druggability.

Novelty

In this work the first structure kinetic relationship (SKR) is presented. Particular polar atoms, ABPAs, are found to act as putative kinetic traps for solvent molecules. Furthermore it is shown that targeting ABPAs can confer kinetic stability in drug design. This can be particularly helpful in lead-optimisation for longer residence times.

Limitations

The study collects a large variety of information from experimental and theoretical sources. Despite this fact it is found that an ABPA is a necessary, but not sufficient condition to act as a kinetic trap. Further properties have to be assessed to refine the theory. Furthermore, the results presented, only show evidence for solvent-protein and protein-ligand interactions. It is likely that the principle is generally applicable to all types of molecular interactions (protein-protein especially) that occur in aqueous environment.

Bibliography

Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25): 10184–9, June 2011.

Robert A Copeland, David L Pompliano, and Thomas D Meek. Drug-target residence time and its implications for lead optimization. *Nature reviews. Drug discovery*, 5 (9):730–9, September 2006.

R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D. E. Shaw. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, pages 1104614108–, July 2011.

L Englert, A Biela, M Zayed, A Heine, D Hangauer, and G Klebe. Displacement of disordered water molecules from hydrophobic pocket creates enthalpic signature: binding of phosphonamidate to the S_1'-pocket of thermolysin. *Biochimica et biophysica acta*, 1800(11):1192–202, November 2010.

Geoffrey A Holdgate and Walter H J Ward. Measurements of binding thermodynamics in drug discovery. *Drug discovery today*, 10(22):1543–50, November 2005.

Gerard J Kleywegt, Mark R Harris, Jin Yu Zou, Thomas C Taylor, Anders Wählby, and T Alwyn Jones. The Uppsala Electron-Density Server. *Acta crystallographica. Section D, Biological crystallography*, 60(Pt 12 Pt 1):2240–9, December 2004.

Jose E. Leysen and Walter Gommeren. Drug-receptor dissociation time, new tool for drug research: Receptor binding affinity and drug-receptor dissociation profiles of serotonin-S2, Dopamine-D2, histamine-H1 antagonists, and opiates. *Drug Development Research*, 8(1-4):119–131, May 1986.

Sten Ohlson. Designing transient binding drugs: a new concept for drug discovery. *Drug discovery today*, 13(9-10):433–9, May 2008.

Yibing Shan, Eric T Kim, Michael P Eastwood, Ron O Dror, Markus A Seeliger, and David E Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–3, June 2011.

David C Swinney. The role of binding kinetics in therapeutically useful drug action. *Current opinion in drug discovery development*, 12(1):31–39, 2009.

David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews. Drug discovery*, 10(7):507–19, January 2011.

Rumin Zhang and Frederick Monsma. The importance of drug-target residence time. *Current opinion in drug discovery & development*, 12(4):488–96, July 2009.

Puentes de Hidrógeno Protegidos: Determinantes Estructurales de la Cinética de Unión. Aplicación en el Diseño de Fármacos.

Peter Schmidtke, F. Javier Luque, James B. Murray, Xavier Barril
en revision, Journal of the American Chemical Society

El control a escala temporal de las interacciones moleculares es una parte esencial de los sistemas bioquímicos, pero muy poco se conoce sobre los factores estructurales que gobiernan la cinética de los reconocimientos moleculares. En el diseño de fármacos, el tiempo de vida de los complejos diana, es el mayor determinante de los efectos farmacológicos, pero la ausencia de relaciones estructurales y cinéticas, impide la optimización racional de esta particularidad. Aquí mostramos que los átomos polares que están casi enterrados en la proteína –una característica común en los lugares de unión de la proteína- tienden a formar puentes de hidrógeno que están protegidos del agua. La formación y posterior ruptura de este tipo de puentes de hidrógeno, conlleva un estado de transición energéticamente penalizado, puesto que esto ocurre asincrónicamente con la hidratación/deshidratación; en consecuencia, los puentes de hidrógeno protegidos del agua se intercambian a muy baja frecuencia. La ocurrencia de este fenómeno se puede anticipar por simple análisis estructural, proporcionando una herramienta novedosa para interpretar y predecir relaciones cinético/estructurales. La validez de este principio se investigó en dos pares de inhibidores de Hsp90, para los cuáles se han determinado datos termodinámicos y cinéticos de manera detallada. El acuerdo entre observaciones macroscópicas y simulaciones moleculares, confirman el papel de los puentes de hidrógeno blindados como trampas cinéticas e ilustran cómo nuestro descubrimiento puede ser utilizado como ayuda en el descubrimiento de fármacos basado en estructuras.

Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics. Application in Drug Design.

Peter Schmidtke[1,2], F. Javier Luque[1,2], James B. Murray[3], Xavier Barril[1,2,4*]

[1]Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Av. Joan XXIII s/n, 08028 Barcelona, Spain

[2]Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Spain

[3]Vernalis (R&D) Ltd., Granta Park, Great Abington, Cambridge CB21 6GB, United Kingdom

[4]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

* To whom correspondence should be addressed: xbarril@ub.edu

Abstract

Timescale control of molecular interactions is an essential part of biochemical systems, but very little is known about the structural factors governing the kinetics of molecular recognition. In drug design, the lifetime of drug-target complexes is a major determinant of pharmacological effects, but the absence of structure-kinetic relationships precludes rational optimization of this property. Here we show that almost buried polar atoms – a common feature on protein binding sites – tend to form hydrogen bonds that are shielded from water. Formation and rupture of this type of hydrogen bonds involves an energetically penalized transition state because it occurs asynchronously with dehydration/rehydration. In consequence, water-shielded hydrogen bonds are exchanged at slower rates. Occurrence of this phenomenon can be anticipated from simple structural analysis, affording a novel tool to interpret and predict structure-kinetics relationships. The validity of this principle has been investigated on two pairs of Hsp90 inhibitors for which detailed thermodynamic and kinetic data has been experimentally determined. The agreement between macroscopic observables and molecular simulations confirms the role of water-shielded hydrogen bonds as kinetic traps and illustrates how our finding could be used as an aid in structure-based drug discovery.

The structure of macromolecules has been used for decades both to understand fundamental biological processes and as an aid in drug discovery.[1] Owing to the type of available experimental data, structure-activity relationships have mainly focused on the thermodynamic properties of end-states. However, biological processes occur at specific and finely controlled timeframes. The biological activity of drugs is also heavily influenced by the kinetics of the drug-target complex[2]. Modulating the kinetic behavior of complexes is, therefore, of fundamental interest both in protein design and in drug discovery. However, this goal is severely hampered by our limited knowledge about the structural factors mediating the association and dissociation processes.

Electrostatic steering has been described as a factor that can speed up the association rate, while leaving the dissociation rate unchanged. This has been used in the design of tighter protein-protein pairs[3, 4], but the concept has limited applicability in drug design as the introduction of charged centers has multiple and important off-target consequences. Coupling association/dissociation to a slower process is another known mechanism to modulate the kinetics of binding. This is illustrated by the DFG-out inhibitors of p38 MAP kinase: as they require a displacement of the activation loop, their on- and off-rates are much slower than those of DFG-in inhibitors[5]. This knowledge is often exploited in the design of kinase inhibitors, but cannot be applied to other target families, as information about conformational transitions and the timescale in which they occur is rare and difficult to obtain. In this paper we demonstrate that formation of water-shielded hydrogen bonds between a ligand and its receptor protein is a viable strategy to increase the kinetic stability of complexes.

The notion of water shielding as a stabilizing mechanism has important precedents in the literature. In fact, water has been touted as the "lubricant of life"[6] and numerous accounts indicate that water acts as a facilitator of motion in proteins and nucleic acids[7, 8]. Similarly, computational studies replicating single-molecule force spectroscopy experiments have long noted that the mechanical stability provided by hydrogen bonds depends on their degree of solvent exposure (see [9] and references therein). These studies clearly point out that, *per se*, hydrogen bonds are stiff structures and the presence of water is necessary to achieve the dynamic exchange characteristic of biological systems. From that point of view, controlling water

accessibility seems a straightforward mechanism to set the pace of events. At the bottom of deep cavities, water may be completely removed, leading to large penalties and extremely slow exchange rates. This is the case of the biotin-streptavidin system, one of the tightest and longest-lived protein-ligand complexes. The interacting pair forms a dry network of hydrogen bonds in the deepest part of the pocket and water entrance through an access channel is thought to be the first event of the dissociation process[10]. This is consistent with the fact that streptavidin mutants that increase the water contents around the hydrogen bond network not only produce a significant loss of potency, but also a large increase in the on- and off-rates[11, 12]. Quite unexpectedly, we find that this effect is also reproduced – at a smaller scale – on solvent exposed areas of the protein surface. Firstly, a description of the phenomena on a test system is presented. Evidence of its occurrence in biological systems is then sought using crystallographic data and molecular dynamics simulations. Finally, the relevance of the principle for drug design is demonstrated on Hsp90 inhibitors.

RESULTS AND DISCUSSION

Dissociation of shielded hydrogen bonds involves a transition state.

In a recent effort to predict the druggability of putative binding sites, we noticed that polar atoms in drug binding sites are located in predominantly apolar environments and tend to be poorly solvent exposed. Yet, they are available for interactions[13]. Given that burial of polar surface area involves a substantial desolvation cost, a functional role for such almost buried polar atoms (hereafter referred as ABPAs) can be assumed. From a thermodynamic perspective, protecting hydrogen bonds from water results in a decreased dielectric constant and subsequent stabilization of the electrostatic interaction[14]. Recently, this effect has been quantified in proteins, demonstrating that hydrogen bonds can be up to 1.2 kcal/mol stronger in hydrophobic environments.[15] Considering that electrostatic effects can be relatively long-range, we were curious to know whether ABPAs could also be related to other fundamental aspects of binding not strictly related to the energetics of the bound state. To that end, we investigated how the level of exposure of a polar atom on the receptor affects the interaction with a ligand along the association pathway. In order to make the problem

tractable and to disentangle the effect of burial from the many other interactions that occur in a real system, we designed a virtual binding site composed of a hydrogen bond acceptor (acetonitrile) surrounded by methane molecules arranged as a half sphere. Although not intended to replicate real biological systems, similar model systems have proved useful to investigate fundamental molecular phenomena [16, 17]. In particular, we wanted to understand if and how changes in the solvent accessible surface area (SASA) of the polar atom and in the local curvature of the receptor could affect the free energy profile of hydrogen-bond formation. To that end, we generated a set of artificial systems that cover a range of values for the parameters A_0 (SASA obtained with a probe of 1.4Å radius) and ΔA (change of SASA as the radius of the probe increases). Full details are provided as Supporting Information. Each one of these systems was then solvated with TIP3P water molecules and multiple steered molecular dynamics (SMD) simulations were used to study the formation of a hydrogen bond between acetonitrile (receptor) and an ammonia molecule (ligand). The corresponding free energy profile was computed using the Jarzynski relationship[18] and is shown on Figure 1A for an exposed donor on a flat surface (black line) and an ABPA on a convex surface (red line). In both cases the formation the hydrogen bond is unfavorable, reflecting the characteristics of the system[19], but it should be noted that we are not interested on the specific values but on the effect that the local environment has on them. In accordance with the above-mentioned dielectric effect, we see a slight tendency to lower the free energy of the bound state as the acceptor group becomes shielded from bulk solvent. However, a more noticeable effect is the appearance of a free energy peak when the ligand moves from the second solvation shell to form a direct contact (Fig. 1a). If the polar atom of the receptor is solvent exposed, exchange of hydrogen bonding partners occurs in a concerted-like manner (water molecule leaves as ammonia approaches) and no transition state is involved (Fig. 1b and black-framed picture in Fig. 1a). However, more restricted environments (small A_0 and negative ΔA values) impose a steric impediment on the exchange, and the water molecule must start to dissociate before the polar atom on the receptor becomes accessible to the incoming ammonia molecule. This results in a situation in which both water and ammonia are removed from the acceptor atom, thus explaining the energetically penalized transition sate (Fig. 1c and red-framed picture in Fig. 1a). As the energy of a hydrogen bond is sharply dependent on the distance,

any uncompensated lengthening of the interaction leads to significant penalties that should be enthalpic in origin. To confirm the expected thermodynamic signature, the free energy profiles of two of the systems were obtained at 7 different temperatures, ranging from 280K to 340K, and a van't Hoff plot was used to obtain the thermodynamic components of the transition state. As shown in Fig. S1 (Supplementary Information), the transition state is due to an enthalpic term that is offset by a favorable entropic contribution, possibly due to a disruption of the solvation shell.

In summary, simulations with the model systems suggest that polar atoms may act as kinetic traps for any interacting partner if they have low SASA (A_0<10Å^2) and are placed in concave local environments ($\Delta A<0$), due to decoupling of the association and hydration processes. Although extrapolation to real systems must be done with care, it seems likely that this effect may also occur in protein binding sites because they are concave and have a high proportion of ABPAs. Naturally, the height of the barrier will depend very much on the particular characteristics of the local environment (flexibility, polarity, shape, etc.) and the values obtained with the model systems should not be taken as reference. However, as the effect stems from a steric impediment on the transition between the bound and unbound states, it is only logical that it will increase with the size of the ligand (Fig. 1d). It should also be noted that even small increases in the height of the transition state might have major effects on the rate constant because they are related by an exponential term[20].

Distribution and mobility of water molecules in crystal structures and MD simulations.

To find evidence of the role of ABPAs as kinetic traps on real biological systems, we investigated protein-water complexes, taking advantage of the wealth of crystallographic data available for such interactions in the protein data bank (PDB)[21]. The presence of an energetic penalty for the exchange between the first and second layers of solvation should have measurable consequences both on the distribution and the dynamics of bound water molecules. Fig. 2a shows the radial distribution function of water molecules around carbonylic oxygens of the protein backbone, derived from a set of 2,704,956 pairs from 5,664 non-redundant crystallographic structures.

Overall, the first, second and third layers of solvation can be identified at 2.5, 5.0 and 7.5Å. Splitting the data according to the surface area of the carbonyl oxygen, it becomes apparent that the probability of having water molecules in intermediate positions between the first and second layer is much lower for ABPAs than for exposed polar atoms, which is consistent with the presence of a transition state in a higher proportion of cases, as expected.

Further evidence of this effect was obtained comparing the spread of the electron density corresponding to water molecules hydrogen-bonded to polar atoms. If a transition state exists, waters will vibrate with shorter amplitude and exchange less frequently, both of which effects will result in a more localized electron density and, in consequence, lower B-factors. As the absolute value of B-factors is often meaningless, we have normalized the B-factor of crystallographic waters relative to the B-factor of the polar atom to which they are attached. The histograms in Fig. 2b, show that water molecules hydrogen-bonded to ABPAs tend to be less mobile than those bonded to more solvent-exposed atoms. Direct measurement of the electronic dispersion offers the same conclusion (Fig. S2). Although not a direct proof of concept, these results are in very satisfactory agreement with the idea that ABPAs are more likely to act as kinetic traps.

In order to obtain a more direct insight into the kinetic behavior of protein-water interactions, an in-house collection of long MD trajectories (≥50ns) corresponding to several unrelated proteins was analyzed. For each carbonylic oxygen on the protein backbone, all water molecules forming hydrogen bonds along the trajectory were identified and the corresponding residence times were calculated. The histogram in Fig. 3 shows that long residence times are exclusive of ABPAs, but most ABPAs do not have a significant effect on the dynamics of water. In agreement with the results obtained with artificial systems, the inset in Fig. 3 also shows that flat or concave environments are a necessary condition for long residence times.

Elucidating the effect of water-shielded hydrogen bonds on protein-ligand complexes.

As the exchange of water molecules between the protein surface and bulk solvent is energetically neutral, the effects described in the previous section are directly linked to the free energy of transition state between the bound and unbound forms. For protein-ligand complexes, the situation is far more complex because formation of a hydrogen bond, or changes in its level of shielding, will have thermodynamic consequences that may mask the kinetic effect. Furthermore, for drug-sized ligands, breakage and formation of the specific hydrogen bonds is only one of many steps in the association and dissociation pathways, which will also include conformational changes of the ligand and the protein as well as formation of transient interactions. Unless it affects the rate-limiting step, the kinetic effect of water shielding will be smeared and hardly noticeable on the macroscopic rate constants. On the other hand, as the level of shielding also depends on the bulk of the ligand molecule, it seems reasonable to expect that the kinetic effect becomes more obvious than in the exchange of water molecules.

In order to understand the impact of water shielding on the binding kinetics of drug-like compounds, we have obtained kinetic, thermodynamic and structural data for several Hsp90 inhibitors. The compounds belong to the resorcinol series, exemplified by NVP-AUY922 [22] (currently in Phase I and II clinical trials for hematologic malignancies and solid tumors) and their study in this context is particularly relevant because slow exchange rates correlate with better cell and *in vivo* activities [23]. Two pairs of compounds have been investigated (**1-4**; Fig. 4a). They all share a common binding mode (PDB entry 2BSM) but differ in their ability to form hydrogen bonds with the carbonyl oxygen of G97, an ABPA ($A_0 \approx 4Å^2$ and $\Delta A < 0$) located at the periphery of the binding site (Fig. 4b). Compound **1** (VER37655) has a methyl in position R_2 and is unable to interact with this atom, whereas the amide moiety in compound **2** (VER49181) forms a hydrogen bond that is shielded from solvent both by the protein and by the rest of the ligand. The difference between compounds **3** (VER49008) and **4** (VER49009) is much more subtle, as they both form a hydrogen bond with G97, but the change of methyl to ethyl alters the level of water shielding.

For the sake of clarity, we will start describing the latter pair, where the kinetic effect of water shielding is more straightforward.

Hsp90 inhibitors: consequences of modifying the level of shielding.

The hydrogen bond between the amide of compound **3** and the carbonyl of G97 is protected from water attack, because there is only one position in the vicinity from where water molecules can access. Furthermore, the density of water at that position is below the normal value (Fig. 5a). Shielding is further increased in compound **4**. Although the ethyl group is free to rotate, MD simulations indicate a clear preference of the group to occupy the position previously taken up by water (Fig. 5b). As expected for such a small change at the periphery of the binding site, the impact is modest. In fact, the binding affinity of both compounds is indistinguishable by SPR (Table 1) as well as in binding and functional assays[24]. ITC experiments, however, reveal a small gain in binding free energy for compound **4** (0.3 kcal/mol). The methyl to ethyl change also has measurable kinetic consequences: the dissociation rate of compound **4** is 1.6-fold slower than **3** (extending the dissociative half-life of the complex[2] from 63 to 100 s). The change in association rate constant, on the other hand, is within the experimental error. Most noticeable, the rise of the dissociation transition state ($\Delta\Delta G^* = 0.3$ kcal/mol) is due to an enthalpic penalty ($\Delta\Delta H^* = 0.6$ kcal/mol) partially offset by an entropic compensation, a thermodynamic signature in line the with the hydrogen bond shielding effect (Fig. S1).

In order to obtain a more direct relationship between structural effects and the macroscopic constants, we simulated the early dissociation process by means of SMD. As shown in Fig. 5, the free energy profile of all compounds starts with a steepest phase, corresponding to rupture of hydrogen bonds, followed by a less pronounced phase in which protein and ligand still form a substantial number of contacts (Fig. S3). The slope during the first stage of the dissociation free energy profile is not as steep for **3** as it is for **4**, leading to a gap of approximately 1 kcal/mol (upper part of Fig. 5a). This indicates that the native hydrogen bonds can be more easily broken in the case of **3**, which is consistent with the larger solvent accessibility of G97 when this compound is bound (lower part of Fig. 5a).

Considering together: i) that the slower compound is the one better protected from water; ii) the enthalpic origin of the transition state; and, iii) the qualitative agreement between macroscopic constants and molecular simulations, it seems logical to assume that the water shielding effect is behind the change in kinetics. This example also demonstrates that water shielding can have a larger effect on dissociation than association, thus resulting in tighter complexes. This can be explained because the rupture of hydrogen bonds is the first dissociation step and is more likely to limit the rate of the process. Conversely, the association rate is more easily influenced by the formation of the encounter complex (hence the effect of electrostatic steering on k_{on} but not k_{off}). It should be noted that, although relatively small, the 60% increase in kinetic stability of ethyl-amides appears to confer better biological activities[24] and the moiety was preserved throughout the lead optimization stage that yielded the clinical candidate[22, 23].

Hsp90 inhibitors: consequences of forming a shielded hydrogen bond.

Calorimetric data (Table 1) shows that **2** binds 1.3 kcal/mol more favorably than **1**, clearly indicating that hydrogen bond formation with G97 is very exothermic. The complex is also kinetically more stable (half-life 3.5 fold longer) due to an increase of 0.7 kcal/mol in the free energy of the dissociation transition state ($\Delta\Delta G^*_{off}$). Qualitatively, this is what would be expected for the formation of an additional water-shielded hydrogen bond. However, the enthalpic component is 5.8 kcal/mol larger in compound **1**, which is in disagreement with the expected thermodynamic signature (Fig. S1). To better understand these contradictory results, we have simulated the early dissociation process by means of SMD. As shown in Fig. 4b, both compounds have a similar slope in the first part of the dissociation free energy profile. But in the case of **1**, the first phase ends when the ligand has moved 1Å away from the ideal position, while this occurs 0.6 Å further away in the case of **2**, coinciding with the first contacts between G97 and water molecules (lower part of Fig. 4b). This demonstrates that the additional water-shielded hydrogen bond opposes dissociation strongly and is the last native hydrogen bond to be broken. At the same time, the fact that the change of slope occurs at different points of the reaction coordinate, suggests that dissociation proceeds through different pathways, which would justify the unexpected thermodynamic signature of the dissociation transition state. Very

satisfactorily, SMD simulations at three different temperatures, followed by van't Hoff analysis, also find that the dissociation of **1** has a larger enthalpic component than **2** (Fig. S4). Monitoring of protein-ligand distances confirms that dissociation proceeds differently for both compounds: the additional hydrogen bond in **2** acts as a pivotal point, delaying dissociation of the pyrazole ring and causing an earlier rupture of the hydrogen bond between the resorcinol and D93 (Fig. S5). These results highlight the validity of the principle in protein-ligand complexes, but also the difficulty to extricate the kinetic effect of water shielding from other kinetic and thermodynamic effects at the macroscopic level for such complex systems.

DISCUSSION

We have demonstrated that water-shielded hydrogen bonds provide kinetic stability by the simple mechanism of decoupling ligand association from water dissociation (and *vice versa*). Several recent but unrelated publications have also found relationships between binding kinetics and hydration. For instance, Liu et al. have demonstrated that a dehydrated state (gas phase) provides kinetic stability to a protein-ligand complex, while hydration stabilizes the transition state[25]. In an impressive study carried out at D. E. Shaw Research, it was found that removal of the last solvation layer was a kinetic barrier in the association process of kinase inhibitors[26]. Researchers at Pfizer found a correlation between displacement of tightly bound water molecules and second-order acylation rate constants of β-lactam antibiotics[27]. We are hopeful that our work provides a theoretical framework to interpret those results and to help uncover structure-kinetic relationships in protein-ligand complexes and any molecular interaction occurring in aqueous environments.

Considering the relationship between kinetic stability of drug-target complexes and bioactivity[2], we have also investigated the possibility of exploiting the water-shielding principle in drug design. Predicting if – and to what extend – a potential hydrogen bond will be shielded from solvent is a trivial exercise using structure-based drug design methods. Comparison of compounds **3** and **4** demonstrates that subtle structure-kinetic relationships can be predicted by simple visual inspection when the water shielding effect is considered. Although in some cases it may be difficult to disentangle the water shielding effect from other consequences of the chemical modification (e.g. **1** vs. **2**), this principle provides two simple recommendations to

improve the kinetic stability of protein-ligand complexes: 1) increase the shielding of existing hydrogen bonds and, 2) when reaching for new interaction, prioritize ABPAs. In an extremely suggestive contribution, Colizzi et al. have shown that true binding modes and true ligands are more difficult to remove from the binding site than decoys.[28] If this is confirmed as a general rule, hydrogen bond shielding and other structural determinants of binding kinetics should also prove useful to develop new scoring functions.

METHODS

Molecular simulations. All molecular simulations were carried out with the AMBER 9 package [29] and the Amber 99 forcefield. Full details are provided as Supplementary Methods.

Analysis of crystallographic structures: Electron density and B-factors. First the PDB was scanned for crystal structures having a resolution better than 2.0 Å and an Rfree factor below 0.3. From this set only proteins with a maximum of 90% sequence similarity were retained. Of this set, 1,723 structures having the electron density maps available in the Uppsala Electron Density Server[30] were considered for this study. Water molecules were considered if they made contacts only with the asymmetric unit (i.e. they were more than 4 Å away from all symmetric units) and had 100% occupancy. For the sake of simplicity, only the water molecules in contact with a single polar atom of the protein were further analyzed (i.e. polar atom-water distance < 3.5 Å and no other polar atom within a 4 Å range of the water coordinates). For those waters, the average and standard deviation of electron density was extracted from a box of 9 grid points around the closest grid point to the water molecule, called ED_{wat}. For the polar atom on the protein, the electron density was extracted based on the same principles as for waters resulting in a quantity called ED_{prot}. Similarly to the derivation of the electron density, the B factors were extracted for the water and the polar atoms, called BF_{wat} and BF_{prot} respectively. As electron densities and B factors are two measures that are not directly comparable between two different crystal structures, we normalized a water-protein atom interaction by calculating the ratio between ED_{wat} and ED_{prot}, as well as BF_{wat} and BF_{prot}, called ED_{ratio} and BF_{ratio}

respectively. Last, the ED_{ratio} was divided in two categories (i) localized electron density and (ii) dispersed density. This division was done per structure using the maximum 30% of ED_{ratio} as definition of localized ED and the minimum 30% of ED_{ratio} as dispersed ED.

Analysis of crystallographic structures: Radial distribution functions. In order to derive the radial distribution of waters around carboxylic oxygens, the previously established dataset was used with the difference that the availability of an electron density map was not necessary. Thus, a total of 5,664 structures were analyzed, resulting in 2,704,956 water-carbonylic oxygen pairs. A continuous form of the radial distribution function [31] was used:

$$g(r) = A(r)^{-1} \sum_{i=1}^{N_{prot}} \sum_{j=1}^{N_{wat}} e^{-C(r-r_{ij})^2} \qquad 1$$

where r is the distance measured, A the area of a sphere of radius r, r_{ij} the distance between the protein atom i and the water atom j, and C a constant damping term set to 10 in this study.

Materials. Compounds **1**-**4** were provided by Vernalis and their synthesis has been published previously[24]. Histidine-tagged Hsp90 was produced as previously described [24, 32].

Isothermal tritation calorimetry (ITC). Isothermal Titration Calorimetry. The ITC measurements were performed using an iTC200 instrument (Microcal, GE Healthcare). All experiments were performed with 20 μM protein at 25ºC, 10 mM HEPES pH7.4, 150 mM NaCl, 0.5 mM EDTA, 0.05% Tween-20 and 1% DMSO. All data was fitted to a one site model using the provided software.

Surface Plasmon Resonance. SPR measurements were performed on a BIAcore T100 instrument (BIAcore GE Healthcare). All experiments were performed on Series S NTA chips (certified) according to provider's protocols with 10 mM HEPES pH7.4, 150 mM NaCl, 25 μM EDTA, 0.05% Tween-20 and 1% DMSO as a running buffer. Histidine-tagged Hsp-90 was immobilized on the sensor surface, reference surfaces without immobilized Ni2+ served as controls for non-specific binding and refractive index changes. The sensor surface was regenerated between experiments with 1 M imidazole and 45% DMSO to eliminate any carry-over of protein and /or

analyte. Data processing was performed using BIAevaluation 1.1 software (BIAcore GE Healthcare Bio-SciencesCorp) with global fitting of the concentration series to a single step Steady State affinity model.

Thermodynamic properties derived from SPR data. The K_{on} and K_{off} rates obtained from SPR experiments were used to obtain binding constants and the corresponding binding free energies:

$$K_a = \frac{k_{on}}{k_{off}} \qquad 2$$

$$\Delta G_{bind} = -RT\ln(K_a) \qquad 3$$

Activation free energies for the dissociation process were obtained using Eyring's equation:

$$\Delta G^*_{off} = -RT\ln\left(\frac{k_{off}h}{k_B T}\right) \qquad 4$$

Thermodynamic decomposition of this property was achieved obtaining K_{off} values at 7 different temperatures in the 6ºC to 30ºC range and using the linear form of Eyring's equation:

$$\ln\frac{k_{off}}{T} = -\frac{\Delta H^*_{off}}{RT} + \ln\frac{k_B}{h} + \frac{\Delta S^*_{off}}{R} \qquad 5$$

The fitted linear relationship between $\ln(K_{off}/T)$ and 1/T had an r^2 value of 0.90 for **2** and 0.96 for the other three compounds (Figure S6). Details on the statistical treatment of experimental data are provided in tables S1-S3.

ACKNOWLEDGMENTS

We thank the Barcelona Supercomputing Center and the Red Española de Supercomputación (RES) for access to computational resources. This work was financed by the Ministerio de Educación y Ciencia (Grant SAF2009-08811). PS is funded by the Generalitat de Catalunya. We are indebted with the Vernalis Hsp90

team for contributing reagents, resources and knowledge. We thank Alexey Rak from Sanofi-Aventis for helpful discussion on quality of crystallographic data, Stephanie Perot for advise on statistical analysis and Vincent Le Guilloux for careful proofreading of the manuscript.

REFERENCES

(1) Congreve, M.; Murray, C. W.; Blundell, T. L. *Drug Discov. Today* **2005**, *10*, 895-907.

(2) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. *Nat. Rev. Drug Discov.* **2006**, *5*, 730-739.

(3) Selzer, T.; Albeck, S.; Schreiber, G. *Nat. Struct. Biol.* **2000**, *7*, 537-541.

(4) Schreiber, G.; Shaul, Y.; Gottschalk, K. E. *Methods Mol. Biol.* **2006**, *340*, 235-249.

(5) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. *Nat. Struct. Biol.* **2002**, *9*, 268-272.

(6) Barron, L. D.; Hecht, L.; Wilson, G. *Biochemistry* **1997**, *36*, 13143-13147.

(7) Nakagawa, H.; Joti, Y.; Kitao, A.; Kataoka, M. *Biophys. J.* **2008**, *95*, 2916-2923.

(8) Roh, J. H.; Briber, R. M.; Damjanovic, A.; Thirumalai, D.; Woodson, S. A.; Sokolov, A. P. *Biophys. J.* **2009**, *96*, 2755-2762.

(9) Nilsson, L. M.; Thomas, W. E.; Sokurenko, E. V.; Vogel, V. *Structure* **2008**, *16*, 1047-1058.

(10) Hyre, D. E.; Amon, L. M.; Penzotti, J. E.; Le Trong, I.; Stenkamp, R. E.; Lybrand, T. P.; Stayton, P. S. *Nat. Struct. Biol.* **2002**, *9*, 582-585.

(11) Freitag, S.; Chu, V.; Penzotti, J. E.; Klumb, L. A.; To, R.; Hyre, D.; Le Trong, I.; Lybrand, T. P.; Stenkamp, R. E.; Stayton, P. S. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 8384-8389.

(12) Baugh, L.; Le Trong, I.; Cerutti, D. S.; Gulich, S.; Stayton, P. S.; Stenkamp, R. E.; Lybrand, T. P. *Biochemistry* **2010**, *49*, 4568-4570.

(13) Schmidtke, P.; Barril, X. *J. Med. Chem.* **2010**, *53*, 5858-5867.

(14) Fernandez, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 113-118.

(15) Gao, J.; Bosco, D. A.; Powers, E. T.; Kelly, J. W. *Nat. Struct. Mol. Biol.* **2009**, *16*, 684-690.

(16) Baron, R.; Setny, P.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 12091-12097.

(17) Setny, P.; Baron, R.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 2866-2871.

(18) Jarzynski, C. *Phys. Rev. E. Stat. Nonlin Soft Matter Phys.* **2006**, *73*, 046105.

(19) Nitrile is a weaker hydrogen bond acceptor than water: the BSSE-corrected interaction energy at the MP2/6-311++G(d,p) level for formonitrile-ammonia and formonitrile-water amounts to -1.7 and -3.5 kcal/mol, respectively.

(20) Laidler, K. J.; King, M. C. *J. Phys. Chem.* **1983**, *87*, 2657-2664.

(21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235-242.

(22) Eccles, S. A., et al *Cancer Res.* **2008**, *68*, 2850-2860.

(23) Brough, P. A., et al *J. Med. Chem.* **2008**, *51*, 196-218.

(24) Dymock, B. W.; Barril, X.; Brough, P. A.; Cansfield, J. E.; Massey, A.; McDonald, E.; Hubbard, R. E.; Surgenor, A.; Roughley, S. D.; Webb, P.; Workman, P.; Wright, L.; Drysdale, M. J. *J. Med. Chem.* **2005**, *48*, 4212-4215.

(25) Liu, L.; Michelsen, K.; Kitova, E. N.; Schnier, P. D.; Klassen, J. S. *J. Am. Chem. Soc.* **2010**, *132*, 17658-17660.

(26) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. *J. Am. Chem. Soc.* **2011**, *133*, 9181-9183.

(27) Han, S.; Zaniewski, R. P.; Marr, E. S.; Lacey, B. M.; Tomaras, A. P.; Evdokimov, A.; Miller, J. R.; Shanmugasundaram, V. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 22002-22007.

(28) Colizzi, F.; Perozzo, R.; Scapozza, L.; Recanatini, M.; Cavalli, A. *J. Am. Chem. Soc.* **2010**, *132*, 7361-7371.

(29) Case, D. A., et al *University of California* **2006**, *9*.

(30) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wahlby, A.; Jones, T. A. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 2240-2249.

(31) Gasteiger, J. J.; Engel, T. In *Chemoinformatics: A Textbook;* Wiley-VCH: Weinheim, 2003; .

(32) Wright, L., et al *Chem. Biol.* **2004**, *11*, 775-785.

FIGURES

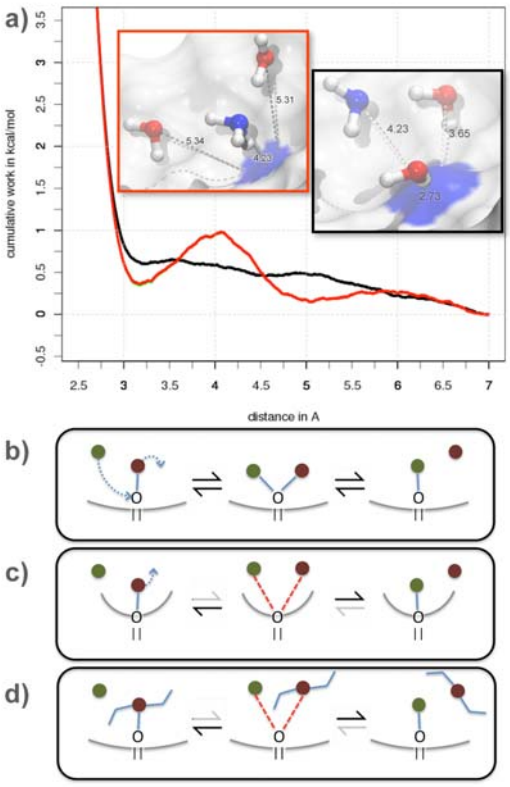


Figure 1. A) Free energy profile of association between a hydrogen bond donor (ammonia) and a model binding site containing a single hydrogen bond acceptor with varying degrees of solvent accessibility. At low levels (A_0=3.3Å^2; ΔA=-1.4; red line) a transition state appears between the bound and unbound states, not present at higher solvent accessibility levels (A_0=12.8Å^2; ΔA=1.3; black line). The insets show the respective configurations at the point in the reaction coordinate where the transition state appears. B) Schematic representation of the water-ligand exchange process with a solvent-exposed polar atom. C) Idem for an almost buried polar atom. D) Same as B, with a bulkier ligand.

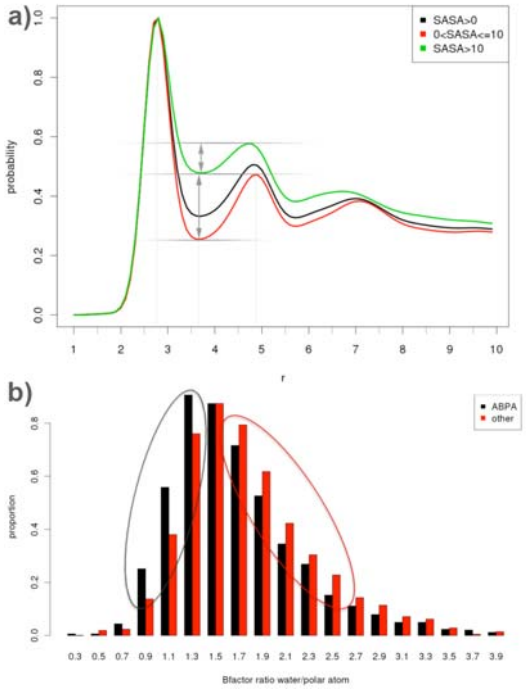


Figure 2. A) Radial distribution function of water molecules around carbonylic oxygens of the protein backbone for a whole set of atom-atom pairs extracted from the PDB (black line). Dividing the set according to the degree of solvent accessibility of the carbonylic oxygen produces a distribution with a deeper minima between the first and second layers of solvation for poorly exposed atoms (red line) and a shallower one for more exposed atoms (green line). B) Histogram showing the distribution of B-factor ratios between interacting water molecules and carbonylic oxygens. The set has been split according to the degree of solvent accessibility of the carbonylic oxygen. Poorly exposed atoms (ABPAs) show a shift towards lower values compared to more exposed atoms, indicating reduced relative mobility of the water molecule.

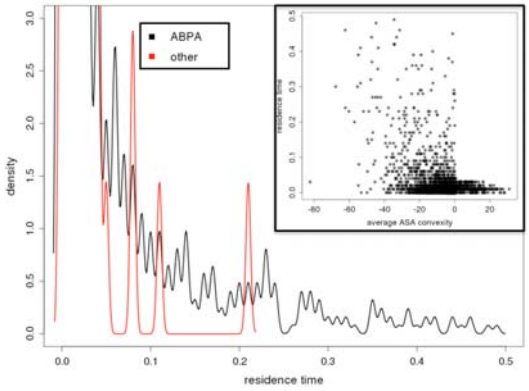
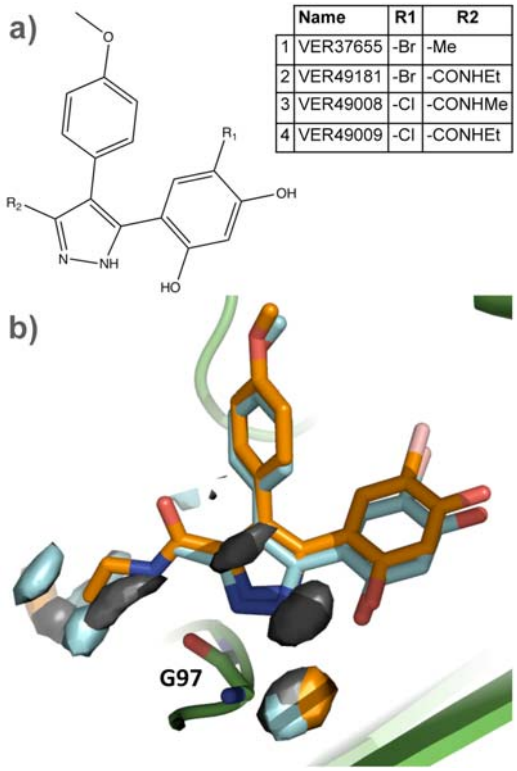


Figure 3. Histogram showing the time length of the interaction between water molecules and the carbonylic oxygen atoms of the protein backbone, obtained from three different molecular dynamics simulations. The distribution for poorly exposed atoms (black line) is shifted towards longer residence times than in the case of exposed atoms (red line). The inset shows a scatter plot of residence times vs. the convexity of the local environment around the carbonylic oxygen, demonstrating that long residence times only occur in concave regions.



	Name	R1	R2
1	VER37655	-Br	-Me
2	VER49181	-Br	-CONHEt
3	VER49008	-Cl	-CONHMe
4	VER49009	-Cl	-CONHEt

Figure 4. A) Chemical structure of four Hsp90 inhibitors. B) Binding mode of Compounds VER37655 (pale blue) and VER49181 (orange) in the ATP-binding site of Hsp90, as obtained from X-ray crystallography. Colored surfaces depict areas where the water density is 3-fold the expected value, as obtained from molecular dynamics simulations. High water density areas are depicted in black for the apo form of the protein, in cyan for the Hsp90-VER37655 complex and orange for the Hsp90-VER49181 complex. Gly 97 is shown in sticks.

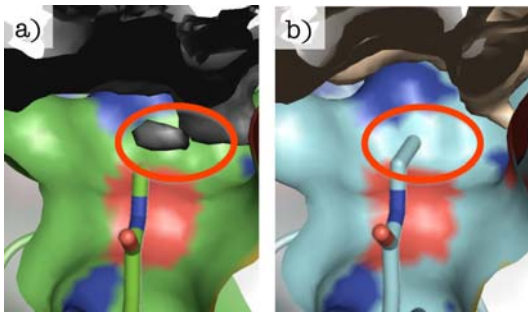


Figure 5. A) Water density isosurface (at 0.75-fold the expected value) for Hsp90-VER49008 complex. B) Idem for Hsp90-VER49009 complex. The red circle highlights the different hydration levels in the vicinity of G97O.

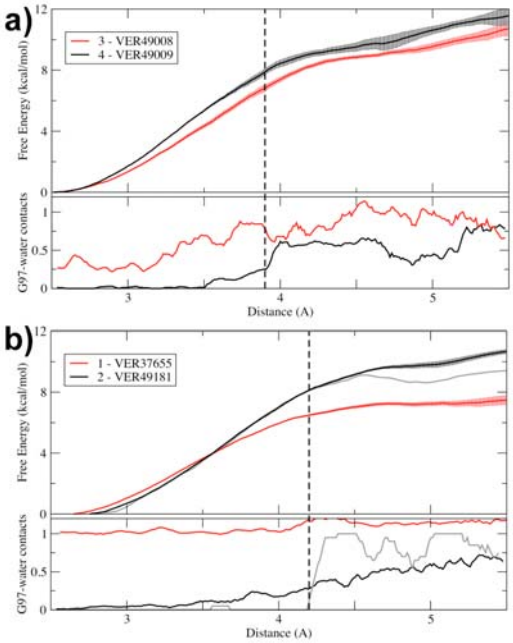


Figure 6. A) Dissociation free energy profile (top) and mean number of G97O-water contacts (bottom) along the dissociation process for VER49008 (**3**) and VER49009 (**4**), derived from multiple SMD simulations. The former has higher solvent exposure of G97O and a less steep free energy profile. B) Dissociation free energy profile of VER37655 (**1**) and VER49181 (**2**) (top) and mean number of water contacts made by G97O along the dissociation process (bottom), as derived from multiple SMD simulations. The thin black line corresponds to an individual SMD simulation of the VER49181-Hsp90 complex. For VER49181, G97O makes no water contacts until the hydrogen bond with the ligand is broken. The rupture of the hydrogen bond occurs on average at point 4.2Å of the reaction coordinate (±0.5Å) and in individual SMD simulations it coincides with a change of slope in the free energy profile (vertical dashed line).

Table 1. Summary of calorimetric (ITC) and Surface Plasmon Resonance (SPR) data. Full statistical details provided in supplementary tables S1-S3.

	ITC (25°)			**SPR (25°)**			**SPR (279K-303K)**		
	ΔG	**ΔH**	**-TΔS**	**K_{on}**	**K_{off}**	**ΔG**	**ΔG^*_{off}**	**ΔH^*_{off}**	**$-T\Delta S^*_{off}$**
1	-9.01	-1.54	-7.47	4.3E+05	2.2E-02	-9.94	19.8	21.4	-1.7
2	-10.32	-2.74	-7.58	4.6E+06	6.1E-03	-12.09	20.4	15.6	4.9
Diff | *(ratio)*	**1.3**[a]	**1.2**[b]	**0.1**[c]	*(0.09)*[a]	*(3.5)*[a]	**2.1**[a]	**-0.7**[a]	**5.8**[a]	**-6.6**[a]
3	-10.13	-3.79	-6.35	1.4E+06	1.1E-02	-11.13	20.3	20.5	-0.2
4	-10.44	-3.69	-6.74	9.9E+05	6.9E-03	-11.10	20.6	21.1	-0.5
Diff | *(ratio)*	**0.3**[b]	**-0.1**[c]	**0.4**[c]	*(1.4)*[c,d]	*(1.6)*[b,d]	**0.0**[c]	**-0.3**[a]	**-0.6**[b]	**0.3**[c]

[a] Significant difference with p-value ≤ 0.01; [b] Significant difference with p-value ≤ 0.05; [c] No significant difference (p-value >0.1); [d] significance calculated from ratio values over the temperature range

SUPPORTING INFORMATION

Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics. Application in Drug Design.

Peter Schmidtke, F. Javier Luque, James B. Murray & Xavier Barril

TABLE OF CONTENTS

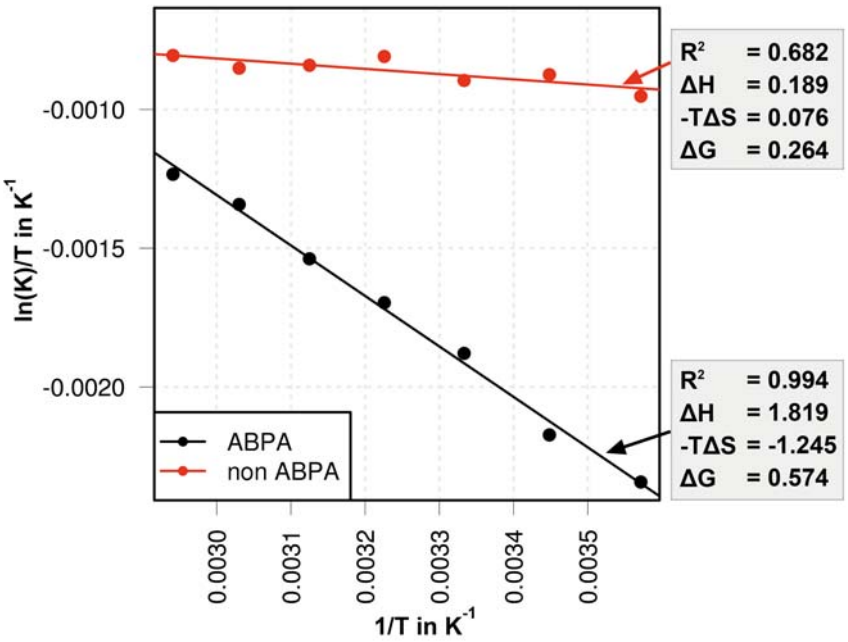


Figure S1. Van't Hoff plot (ln(K)/T vs. 1/T) of model systems containing a solvent exposed polar atom (red line) or an almost buried polar atom (black line). Fitting quality (r^2) and derived thermodynamic properties are also shown.

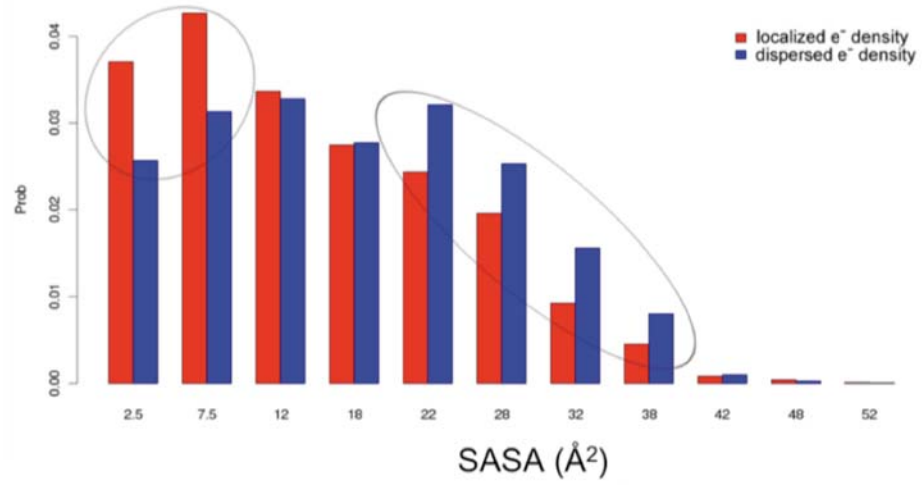


Figure S2. Histogram showing the distribution of 6,813 water molecules with more localized (red) or dispersed (blue) electron density according to the solvent accessible surface area of the carbonylic oxygen of the protein backbone with which they interact. Poorly exposed atoms (ABPAs) have a tendency to bind water molecules with a more localized electron density, whereas well exposed atoms bind a larger proportion of waters with dispersed electron densities.

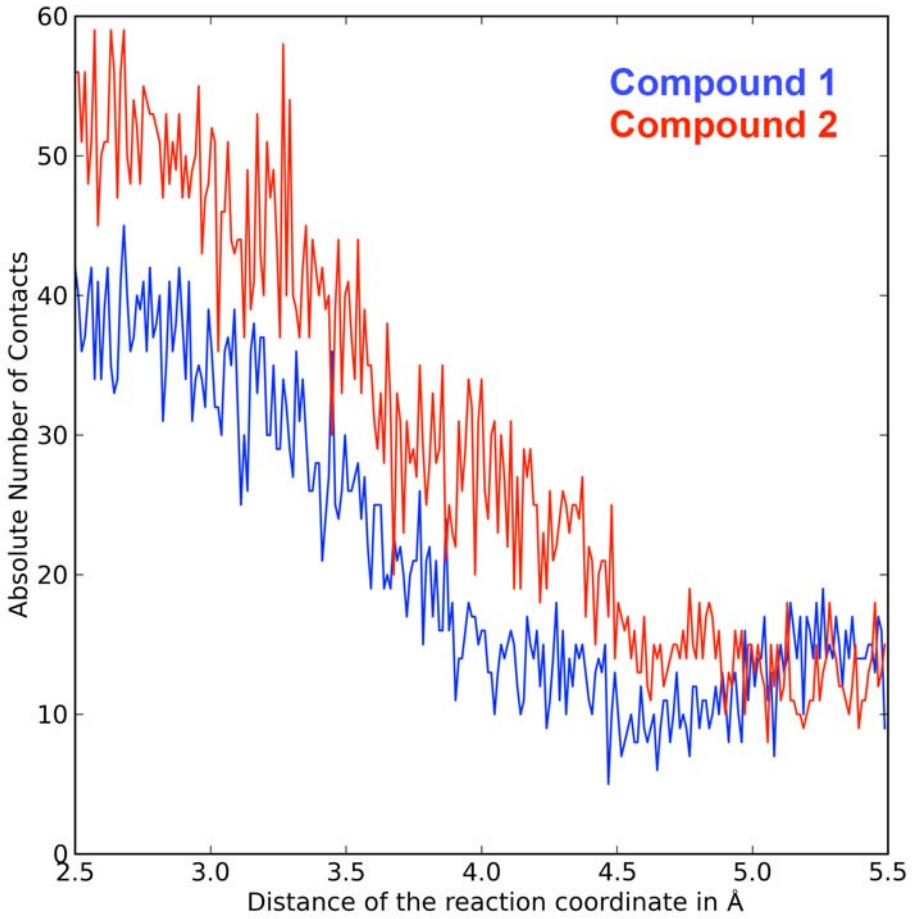


Figure S3. Total number of contacts between ligand **1** (blue) and **2** (red) along the SMD trajectory. A contact is defined as any protein atom closer than 4.0Å from any ligand atom. The ligand has not been pulled completely out of the binding site because the purpose of SMD simulations was to study the early dissociation stage, where hydrogen bonds are broken.

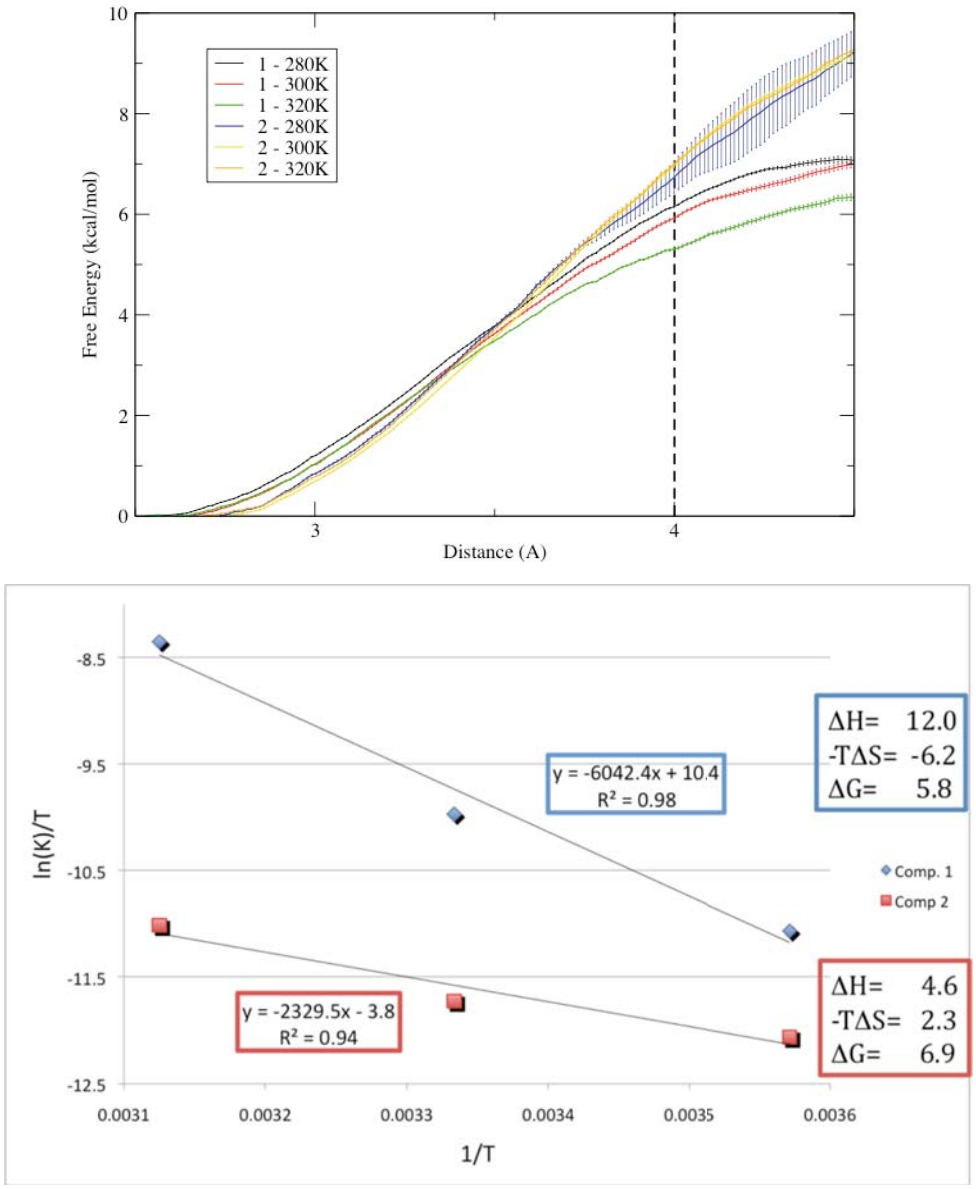


Figure S4. Dissociation free energy profile of **1** and **2** at 3 different temperatures (top). The free energy of dissociation at 4Å (1.5Å away from the equilibrium distance) is approximately constant for **2**, but varies substantially for **1**. van't Hoff analysis of the values at this point (bottom) provides ΔH and -TΔS(300K) values of 12.0kcal/mol and -6.2kcal/mol respectively for **1** (experimental values for the whole dissociation process are 21.4 and -1.7, respectively). For **2**, ΔH=4.6 kcal/mol and -TΔS(300K)=2.3kcal/mol (experimental values for the whole dissociation process are 15.6 and 4.9, respectively). The excellent agreement indicates that the water-shielded hydrogen bond formed by compound **2** does indeed provide additional kinetic stability (Fig. 4), but its thermodynamic signature is lost because dissociation proceeds differently than for compound **1**.

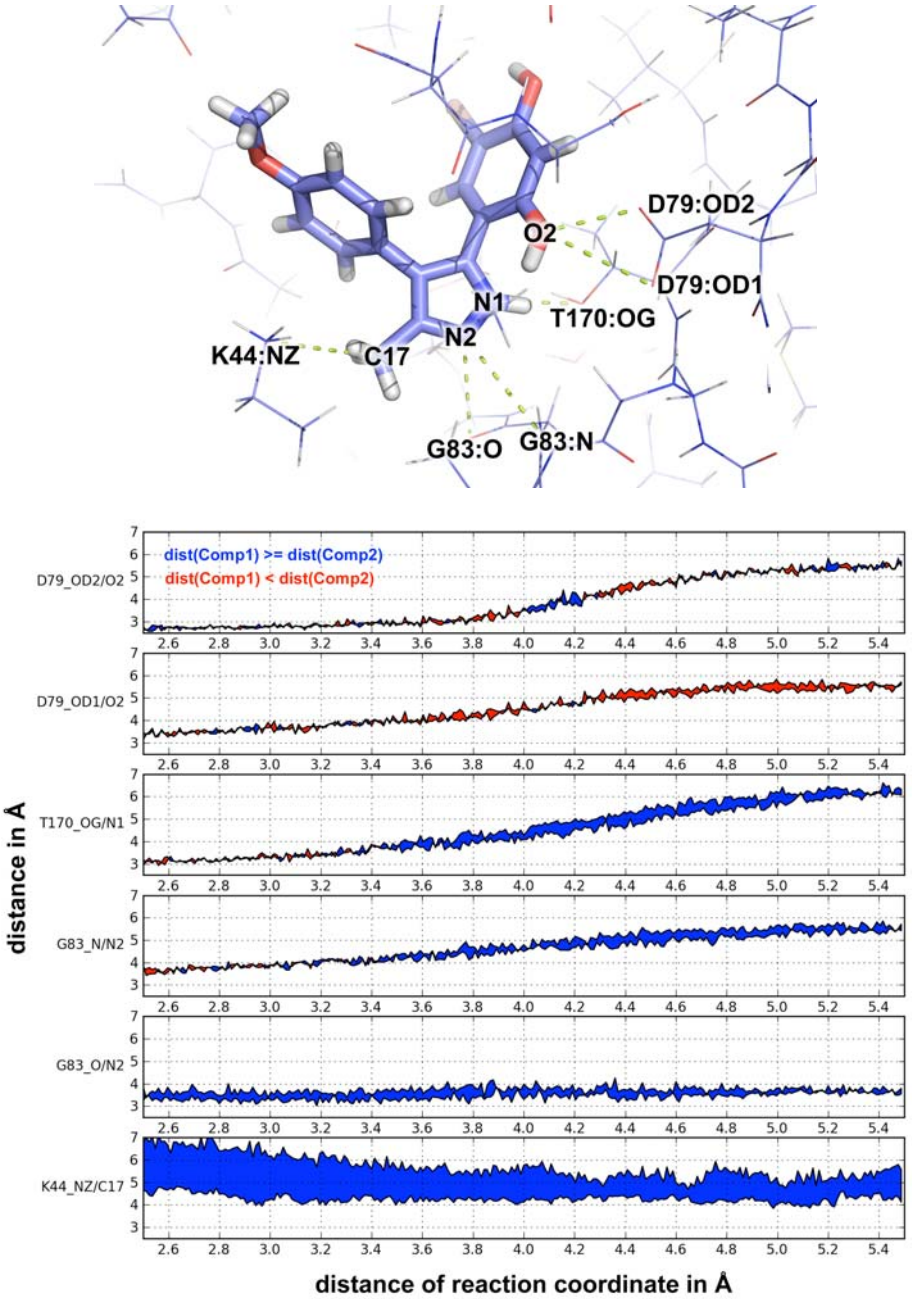


Figure S5. Monitoring of selected distances (shown on top) along the SMD simulations. All values correspond to the Boltzmann-averaged distance of all SMD trajectories (see eq. 7). The plots indicate that **1** separates the pyrazole ring earlier than **2** from the neighboring residues G98 and T184 (labeled as G83 and T170, respectively). The resorcinol ring of **2**, on he other hand, detaches earlier from D93 (labeled as D79) than **1**. Taken together, these results indicate that the additional hydrogen bond of compound **2** acts as a pivotal point, changing the dissociation pathway.

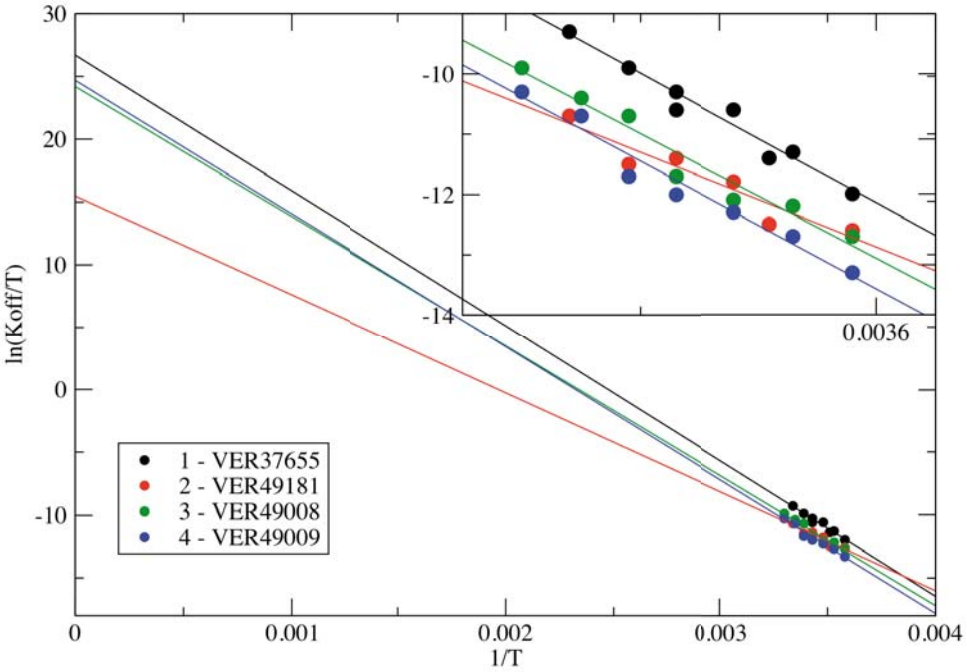


Figure S6. Eyring plot of the dissociation rate constant for compounds 1-4. The inset zooms in the experimentally investigated temperature range (6ºC to 30ºC). Correlation coefficients are 0.90 (red line) and 0.96 (black, blue and green lines).

Table S1. Detailed statistical analysis of ITC data collected at 25ºC. All tests are one-tailed because the goal was to find significant increase/decrease between pairs of compounds. Stoichiometry (N) ranges from 0.89 to 0.92 for all measurements.

	ITC (25º)			
	K_A	ΔG_{bind}	ΔH_{bind}	$-T\Delta S_{bind}$
1 (n=3)	4.3E+06 ± 1.7E+06	-9.01 ± 0.27	-1.54 ± 0.26	-7.47 ± 0.50
2 (n=3)	3.7E+07 ± 2.7E+06	-10.32 ± 0.04	-2.74 ± 0.09	-7.58 ± 0.11
Diff | *(ratio)* P-value (t-test)	*(0.12)* 0.01453	**1.3** 0.008712	**1.2** 0.01178	**0.1** 0.3858
3 (n=3)	2.7E+07 ± 3.9E+06	-10.13 ± 0.08	-3.79 ± 0.34	-6.35 ± 0.39
4 (n=3)	4.5E+07 ± 4.6E+06	-10.44 ± 0.06	-3.69 ± 0.25	-6.74 ± 0.26
Diff | *(ratio)* P-value (t-test)	*(0.61)* 0.0003512	**0.3** 0.001503	**-0.1** 0.1774	**0.4** 0.3967

Table S2. Detailed statistical analysis of SPR data collected at 25ºC. All tests are one-tailed because the goal was to find significant increase/decrease between pairs of compounds . t-test is used to compare compounds **3** and **4** due to the difference size of the samples.

	SPR (25º)		
	K_{on}	K_{off}	ΔG
1 (n=4)	4.3E+05 ± 1.3E+05	2.2E-02 ± 5.1E-03	-9.94 ± 0.30
2 (n=4)	4.6E+06 ± 6.4E+05	6.1E-03 ± 6.7E-04	-12.09 ± 0.15
Diff | *(ratio)* P-value (t-test)	*(0.09)* 0.00068	*(3.5)* 0.00312	**2.1** 0.00041
3 (n=2)	1.4E+06 ± 3.8E+05	1.1E-02 ± 3.3E-03	-11.13 ± 0.07
4 (n=3)	9.9E+05 ± 3.8E+05	6.9E-03 ± 1.3E-03	-11.10 ± 0.20
Diff | *(ratio)* P-value (Welch t-test)	*(1.4)* *0.15690* [b]	*(1.6)* *0.15390* [a]	**0.0** 0.40600

[a] The difference becomes significant (P-value t-test = 0.0396) when considering K_{off} ratios at all temperatures tested (see Table S3). [b] The difference remains not significant (P-value t-test = 0.2973) when considering K_{on} ratios at all temperatures tested (see Table S3).

Table S3. Detailed statistical analysis of SPR data collected over the 279K-303K temperature range. All tests are one-tailed because the goal was to find significant increase/decrease between pairs of compounds.

	SPR (279K-303K)				
	K_{on}	K_{off}	ΔG^*_{off}	ΔH^*_{off} [a]	$-T\Delta S^*_{off}$ [a]
1 (n=8)	-	-	19.8 ± 0.11	21.4 ± 0.45	-1.7 ± 0.46
2 (n=8)	-	-	20.4 ± 0.16	15.6 ± 0.96	4.9 ± 1.00
Diff | *(ratio)*	*(0.13 ± 0.04)*	*(3.40 ± 0.89)*	**-0.7**	**5.8**	**-6.6**
P-value (t-test)	4.3E-10 [b]	0.0024 [b]	1.11E-05	7.67E-08	3.94E-08
3 (n=7)	-	-	20.3 ± 0.12	20.5 ± 0.60	-0.2 ± 0.61
4 (n=7)	-	-	20.6 ± 0.13	21.1 ± 0.67	-0.5 ± 0.71
Diff | *(ratio)*	*(1.38 ± 1.08)*	*(1.64 ± 0.49)*	**-0.3**	**-0.6**	**0.3**
P-value (t-test)	0.2973 [b,c]	0.0396 [b]	0.00163	0.04958	0.21020

[a] multiple ΔH^* and ΔS^* values are obtained leaving one data point out each time and fitting the remaining data to eq. 5. [b] paired one tailed t-test comparing the ratios of rate constants of size n between the two molecules to a random normal distribution of size n with parameters μ=1 and σ taken as the standard deviation of the observed distribution. The process is repeated until the average p-value of the t-test reaches convergence. [c] The lack of significance for such small change (~50% increase) is not unexpected because on-rate measures bear larger uncertainties than off-rate measures.

MATERIALS AND METHODS

Generation of artificial binding site models

Spheres of different radii (5, 7 & 10Å) with uniformly distributed points were generated using a program written by Paul Bourke available in the internet (http://local.wasp.uwa.edu.au/~pbourke/geometry/spherepoints/source1.c). The number of points was chosen to obtain a uniform separation of 3.0Å between neighbour points. The coordinates of one of the points was used to place the nitrogen atom of an acetonitrile molecule (N-point). The first two layers of points around the N-point (distance < 7Å) were retained and their coordinates used to place methane molecules, thus obtaining an artificial binding site consisting of a hydrogen bond acceptor surrounded by a hydrophobic spherical cap of varying curvature. The ΔA parameter is mostly determined by the radii of the sphere, while modifications of the A0 parameter were obtained displacing the acetonitrile molecule along the vector that joins the centre of the sphere with the N-point (in-plane and 1 Å or 2 Å off-plane in each direction, resulting in a total of 15 artificial "binding sites"). The same vector was used to place the ligand molecule (ammonia) 7 Å away from the acetonitrile's N. Atomic charges for ammonia, acetonitrile and methane were obtained using the RESP procedure[5] and atom types were manually assigned to obtain the van der Waals parameters. Using Leap from the Amber 9 package[6], the systems were solvated with TIP3P[7] water molecules and placed in a periodic box extending 10Å further from the solute coordinates.

Steered molecular dynamics of artificial systems.

The hydrogen atoms of the system and water molecules were minimized using the steepest descent algorithm with a maximum number of iterations of 5000, keeping all carbon and nitrogen atoms of the atoms restrained with a weight of 100 kcal/mol/Å^2. The non-bonded terms, which initially were down-weighted to avoid huge steric collisions, were gradually weighted up, allowing a smooth minimisation. During all equilibration and production steps a Langevin thermostat with a collision frequency of 6 was used. A first equilibration keeping the methane half sphere, the acetonitrile and the ammonia restrained using a force constant of 10 kcal/mol/Å^2 was performed during 20ps while heating from 200K to 250K and keeping a constant volume. Next, a second equilibration step of 25ps under constant pressure was performed. During this equilibration the system was heated up until the final 300K. Next a production MD of 1ns was run, keeping the positions of all carbon and nitrogen atoms restrained with a force constant of 10 kcal/mol/Å^2. 20 snapshots equally spaced along the 1ns trajectory were extracted as starting point for the steered MD (SMD) simulation. In each of those 20 simulations, the ammonia was pulled during 1ns from a 7.0Å distance to 2.5Å distance using only the distance as reaction coordinate (5000 kcal/mol/Å^2 as force constant) while tracking the work necessary to perform the displacement. During SMD all atoms of the artificial binding site were constrained, while water molecules were unrestrained, as was the ammonia molecule (except for the reaction coordinate distance). The simulations were run

in the NPT ensemble with isotropic position scaling. The resulting work measured along the reaction coordinated was Boltzmann averaged to derive the change in free energy between the start and end state of the simulation using the Jarzynski relation[8]:

$$e^{\left(\frac{-\Delta G}{k_B T}\right)} = \left\langle e^{\left(\frac{-W}{k_B T}\right)} \right\rangle \qquad 6$$

	R10_100			**R10_080**		
T	ΔG_{min}	ΔG_{max}	$\Delta\Delta G_{TS}$	ΔG_{min}	ΔG_{max}	$\Delta\Delta G_{TS}$
280	0.6341	1.29	0.6559	0.1767	0.4433396	0.2666396
290	0.457	1.0871	0.6301	0.258965	0.5125946	0.2536296
300	0.5847	1.1484	0.5637	0.2687	0.537395	0.268695
310	0.5653	1.09117	0.52587	0.29545	0.5462832	0.2508332
320	0.547	1.0392	0.4922	0.3212	0.59026	0.26906
330	0.55678	0.99971	0.44293	0.3405247	0.6214365	0.2809118
340	0.5007	0.92007	0.41937	0.39503	0.6688633	0.2738333

Table S4. Free energy values (relative to the unbound state) of the bound state (ΔG_{min}), the transition state (ΔG_{max}), and the difference between them ($\Delta\Delta G_{TS}$). R10_080 is an artificial system containing a solvent exposed polar atom, R10_100 is an artificial system containing an ABPA (A0 < 10Å^2; ΔA < 0).

In order to determine the standard enthalpy and entropy change of the system the Van't Hoff equation was used to relate the equilibrium constant with the latter two. For this analysis two configurations of a methane half sphere of 10Å radius were considered. One having an exposed acetonitrile not showing a transition state (N atom protruding 2Å towards the center of the sphere; R10_080) and a second one with the N atom in-line with the methane half sphere (R10_100). For each one of those states, production and SMD simulations were carried out at 7 different temperatures ranging from 280 to 340K. In order to improve statistics, in this case 100 SMD simulations were run at each temperature. Using the Jarzynski relationship again, the free energy profile of the approach of the ammonia towards the acetonitrile was derived and the free energy of the transition state ($\Delta\Delta G_{TS}$) was calculated as the difference between the minimum corresponding to the bound state (ΔG_{min}) and the maximum separating bound from unbound state (ΔG_{max}) (in all cases, $\Delta G_{unbound}$=0). The corresponding values are provided in Table S4. Thermodynamic decomposition is achieved via van't Hoff analysis of the data (plotted in Fig. S1).

Residence time of water molecules

Three 3 MD trajectories of unrelated proteins (HSP90, HDAC8 and PKCΘ) of 78.5ns, 100ns and 60ns, respectively, were used to investigate the residence time of water molecules on the protein surface. The MD trajectories were obtained using the Amber MD package[6] and the Amber 99 forcefield. For each polar atom on the protein the residence time of waters in contact with them was tracked. A water protein atom contact was tracked if the distance between the

oxygen atom of the water and the polar atom of the protein was below 2.8Å. In order to normalize the residence time of waters on each polar atom of the protein the total time a polar atom was in contact with water was divided by the total time of the simulation. This ratio was then further normalized using the total number of waters the polar atom is in contact with during the whole simulation.

Molecular dynamics of Hsp90-inhibitor complexes.

The coordinates of the Hsp90-VER49009 complex (PDB code 2BSM[3]) were used to obtain the starting geometries for all the simulations. The ligand coordinates are taken directly from the PDB file in the case of compound 4; obtaining compounds 1,2 and 3 simply involved atom deletions on the ethyl-amide moiety and/or substitution of the chlorine atom for bromine. Partial charges were obtained using the RESP procedure[5], the rest of parameters are taken from similar atom types in the force-field. Each ligand-protein complex was placed in a truncated octahedral box spanning 13 Å further from the solute in each direction and solvated with TIP3P water models[7]. The average size of the systems is 36,698 atoms. System equilibration involved 1) 2,000 minimization steps, restraining the protein heavy atoms to their original coordinates; 2) 1,000 minimization steps (no restraints); 3) assign initial velocities corresponding to a distribution at 100K and gradually warm up to 300K in 200ps in the NVT ensemble; 4) run 50ps in the NPT ensemble. Finally, MD simulations in the NVT ensemble were carried out for 50ns. Usually, fluctuations of the protein backbone have root mean squared deviations (RMSD) below 2Å relative to the initial crystallographic, indicating that the structure is preserved (Fig. S8a). The RMSD of the inhibitors is in the 1.0Å to 1.5Å region, clearly showing that the binding mode does not change (Fig. S8b). Coordinates and velocities saved along the 10-50ns trajectory range were used as starting points for steered molecular dynamics, thus ensuring the diversity of starting configurations.

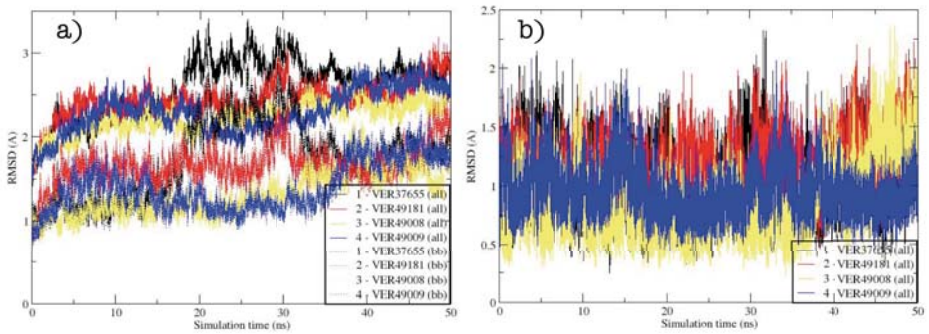


Figure S8. A) RMSD values (best fit to the initial crystallographic structure) for the protein backbone (dashed lines) and all heavy atoms (solid lines). B) RMSD values of the ligand heavy atoms (fitting the protein backbone to the crystallographic structure).

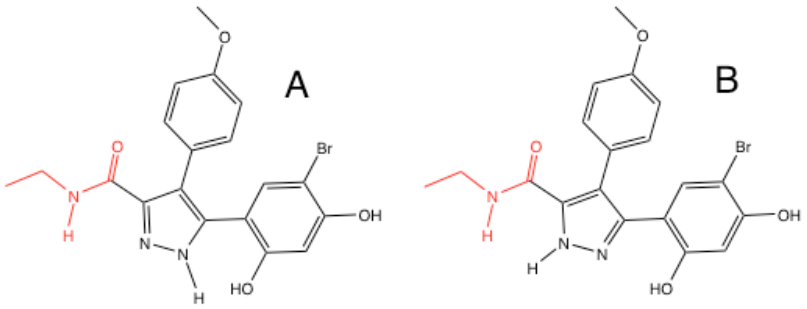


Figure S9. Possible tautomeric forms of compounds **1** (black atoms) and **2** (whole structure).

$\Delta G_{binding}$	Tautomer A	Tautomer B	$\Delta\Delta G_{binding}$
1	**-23.16**	**-25.28**	**2.1**
	-18.25	*-18.89*	*0.6*
2	**-34.08**	**-19.88**	**-14.2**
	-29.15	*-16.88*	*-12.3*

Table S5. MM-PBSA binding free energies. **Bold** is considering three interstitial water molecules as part of the protein, *italics* considering only the protein and the ligand.

As the pyrazole ring of compounds **1**-**4** has two tautomeric forms, both were initially considered for compounds **1** and **2** (Fig. S9). For compound **2**, the internal energy can vary substantially with the orientation of the amide, but only the conformation seen in the crystal structure was considered (i.e. pyrazole N and amide N in *cis*) as the *trans* configuration would involve a repulsion between the O of the amide and G97O. For **1**, the energy of the molecule *in vacuo* is very similar for both tautomers (A favored by 0.8 kcal/mol at the HF/6-31G* level). Compound **2**, on the other hand, favors state A by 6.5 kcal/mol, as tautomer B involves electrostatic repulsion between the N-H groups. For both compounds and tautomers, the binding mode remains stable during the 50ns MD simulations. We carried out MM-PBSA energy decomposition analysis (as implemented in Amber[6]) to estimate the binding free energy (Table S5). Although approximate, the values for the internal energy and the interaction energy with Hsp90 clearly point out that only tautomer A needs being considered for compound **2** and, by extension, for the other amide-bearing compounds (**3** and **4**). In the case of compound **1**, the tautomeric preferences are less clear and both forms were considered in further studies.

Steered Molecular Dynamics of Hsp90-inhibitor complexes.

Steered molecular dynamics (SMD) were used to investigate ligand dissociation. This involves the use of a restrain to pull the ligand away from its binding site. Unlike other systems, were the ligand may use multiple pathways[9], in the case of Hsp90, the choice of reaction coordinate is facilitated by the shape of the binding site, which is located at the bottom of a wide well, clearly indicating the

exit route. Nevertheless, as the choice of reaction coordinate can influence the results, initially we investigated several restraining schemes (Table S6). The chosen reaction coordinate provided the lowest free energy profile amongst those that successfully lead to dissociation. Each system was subjected to multiple SMD simulations, displacing the ligand from its equilibrium distance (2.5Å) to 5.0Å in 0.5 or 1ns.

Name	Equilibrium distance (Å)	Receptor reference point	Ligand reference point
RC1	2.5	Cα of residues L48, N51, S52, A55, D93, G95, I96, G97, M98, F138, T152, T184 & V186	Non-hydrogen atoms of resorcinol and pyrazole rings
RC2	0.5	Cα of residues L48, N51, S52, D54, A55, K58, D93, G95, I96, G97, M98, D102, L107, G108, V136, G137, F138, Y139, V150, T152, H154, T184 & V186	Non-hydrogen atoms of resorcinol and pyrazole rings
RC3	12.0	Cα of residues V150, T184 & V186	Non-hydrogen atoms of amide moiety
RC4	12.5	Cα of residue F138	Non-hydrogen atoms of resorcinol ring
RC5	10.0	Cα of residues L48, N51, I91, F138 & V186	Non-hydrogen atoms of resorcinol ring

Table S6. List of restraining schemes investigated. When multiple atoms are listed, the reference point is their centroid.

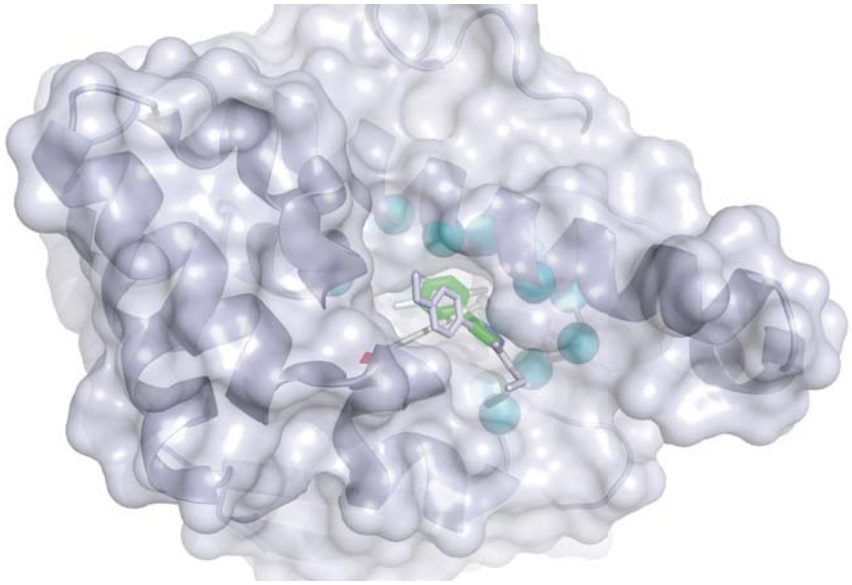


Figure S10. Structure of Hsp90-VER49009 (2BSM) showing the initial direction of the selected reaction coordinate (RC1; arrow with red tip). The cyan spheres correspond to Cα atoms used to define the receptor reference point (small sphere at the arrow begin). The green portion of the ligand is used to define the

ligand reference point (arrow intersection). Note that the pulling direction changes along the simulation.

The dissociation free energy profile was obtained using Jarzynski's relationship (Eq. 6). For each system, the relationship was applied on 500 different combinations of N/2 simulations (where N is the total number of SMD simulations for that system). The final free energy profiles provided in Fig. 4 correspond to the resulting average and standard deviation values. The number of simulations was chosen to ensure sufficiently small uncertainties in the profile. For compound **1**, SMD simulations were carried out for both tautomeric forms but, as they yielded very similar results, only those for state A are shown.

To obtain the water contacts profile, the number of water molecules closer than 3.4 Å from G97_O was counted with ptraj along each simulation, and the values at each snapshot were averaged using the same Boltzmann weights used in Eq. 6:

$$F_i = \frac{e^{\left(\frac{-W_i}{k_B T}\right)}}{\sum_{j=1}^{N} e^{\left(\frac{-W_j}{k_B T}\right)}} \qquad 7$$

$$X_{mean} = \sum_{i=1}^{N} F_i X_i \qquad 8$$

This formalism ensures that the number of water contacts (or any other property X of the system) reflects the average state of the system on that point in the reaction coordinate.

ASSOCIATION FREE ENERGY PROFILES OF AN ARTIFICIAL LIGAND-RECEPTOR COMPLEX

On Fig. S11, the detailed free energy profiles for all the artificial systems analyzed in this work are shown. The profile is obtained from SMD where the ammonia molecule is pulled from 7 Å to 2.5 Å towards the acetonitrile in the bottom of the cavity. Fig. S11a shows the steered MD simulations for a relatively flat artificial cavity (sphere radius = 10Å). One can observe that using a solvent exposed acetonitrile (R10_080) a ground state is present at around 3.2 Å. As the hydrogen bond acceptor becomes less exposed (R10_090 – R10_102.5) the transition becomes more evident, until the polar atom becomes totally buried (R10_105 and R10_110), when it finally disappears. Despite the visible

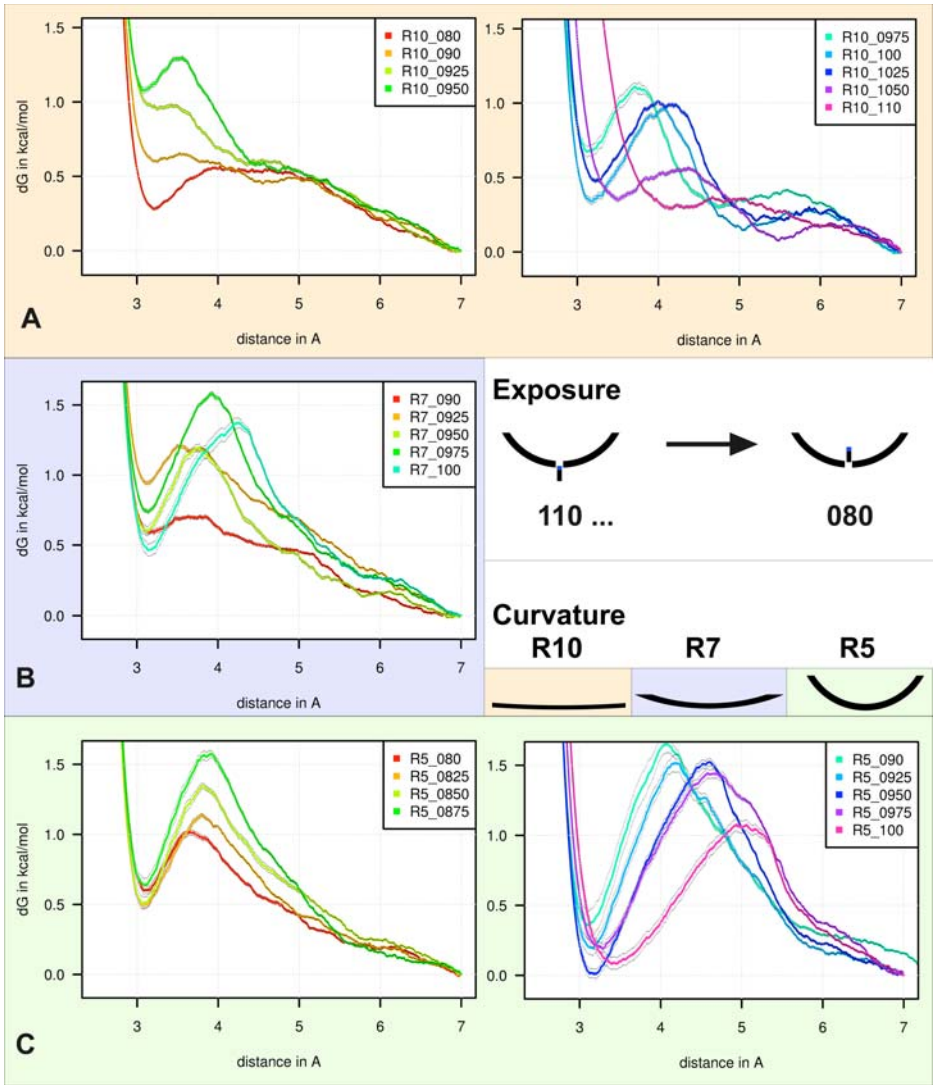


Figure S11. Free energy profiles for the association process between an ammonia molecule and the different model binding sites.

presence of the transition state in those flat pocket systems, the difference of the free energy at the transition state versus the ground state rarely surpasses 0.7 kcal/mol. As can be observed in Fig. S11b (sphere radius = 7Å) and S10c (sphere radius = 5Å), as the curvature of the pocket augments, the height of the transition state also increases, reaching up to 1.5 kcal/mol for R5. This observation indicates that, in more closed cavities, the replacement of the first solvation layer is energetically more costly due to wider decoupling between the dissociation and association processes; i.e. the leaving group has to reach longer distances (and pay a concomitantly higher energetic cost) before exchange occurs. This is apparent comparing the position of the transition state for R10_100 (4Å), R7_100 (4.3Å) and R5_100 (5Å). In summary, higher curvature and shielding from environment increases the height of the transition state (relative to the bound state), but it is noteworthy that, although weaker, this phenomenon is also observed on shallow surface patches like R10 studied here.

The left hand-side of Fig. S12 presents the free energy curves obtained for the artificial system with a 10Å radius with different solvent exposure of the ligand atom (same plots as in Fig. S11). The right hand side shows the solvent accessible surface area (ASA) of the polar atom (A0) and its variation with increasing radii of the probe used to calculate the surface area. From top to bottom of the figure, A0 is lowered until total disappearing. The slope of the ASA profiles (ΔA) is positive initially and becomes negative for low A0 values. The appearance of a transition state (indicated by orange arrows in Fig. S11) can be related to A0 and ΔA using the following simple rules:

1. A transition state is observed when $\Delta A<0$
2. A transition state is observed when $2Å^2 < A0 < 10Å^2$

These rules are also maintained in the artificial systems with smaller radii (R5 and R7). The ASA profiles are extremely fast to calculate, whereas SMD simulations are very costly. Thus, the ASA can be used as a first general filter to estimate if a polar atom and its environment could potentially induce a transition state upon approach of a putative ligand.

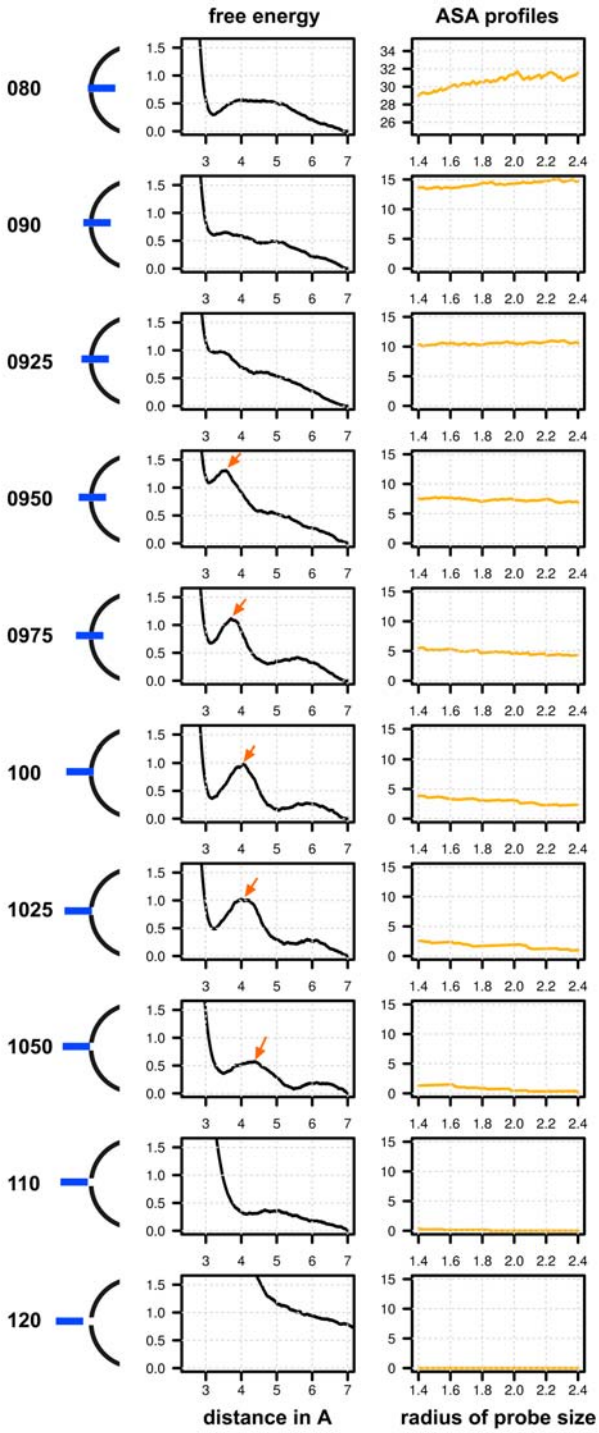


Figure S12. Association free energy profiles and ASA profiles of 10 artificial systems with identical surface curvature and decreasing polar atom exposure.

RELATION WITH THE DEHYDRON THEORY

Discussion of our results with colleagues has often lead to comparison with the so-called dehydron theory of Prof. Ariel Fernandez and co-workers[10, 11]. Both models are absolutely compatible (even complementary) as they focus on different aspects and areas of application. This section aims to clarify similarities and differences between them.

	Dehydron theory	**Kinetic trap theory**
Definitions	Dehydron is an intramolecular hydrogen bond incompletely shielded from water and, in particular, an amide-carbonyl hydrogen bond in the protein backbone incompletely "wrapped" by nonpolar side-chains.	ABPA (almost buried polar atom) is a polar atom with a small solvent accessible surface area (<10Å^2) in a flat or concave surface local area. When engaged in intermolecular interactions, ABPAs tend to form shielded hydrogen bonds.
Consequence of water-shielding a hydrogen bond	Lowered dielectric constant enhances the electrostatic interaction, favoring the bound (folded) state ($\Delta G_{bind}<0$; $\Delta H_{bind}<0$), regardless of the association process.	Decoupling of the water-ligand exchange process. Appearance of a transition state regardless of the overall thermodynamics of the process. For simple systems (e.g. protein-water binding), the expected thermodynamic signature is: $\Delta G^{\ddagger}>0$; $\Delta H^{\ddagger}>0$; $\Delta S^{\ddagger}>0$. Resulting in slower on- and off-rates (binding affinity unchanged). For more complex systems (e.g. drug-protein interactions) the off-rate is more likely to be affected (increased binding affinity) and the thermodynamic signature can become smeared.
Consequence of shielding "defects"	Promote removal of surrounding water (to achieve better ΔG) through protein associations or ligand binding.	Water-ligand exchange can occur in a more coordinated fashion, the transition state decreases, faster on and off-rates.

Application	Dehydrons can be viewed as hydrophobic surfaces. Ligands interacting with dehydrons should have better ΔG_{bind} and may become more selective because dehydrons are not conserved across protein paralogs.	ABPAs should be seen as preferential hydrogen bonding sites if kinetic stability is a desired property. Slower kinetics (and, possibly, better affinity) can also be achieved by increasing the level of shielding of existing protein-ligand hydrogen bonds.
Relation with each other	Not concerned with kinetic effects.	Kinetic effects will usually occur concomitantly with thermodynamic effects.

COMPLETE REFERENCES 22, 23 & 32

[22] S. A. Eccles, A. Massey, F. I. Raynaud, S. Y. Sharp, G. Box, M. Valenti, L. Patterson, A. de Haven Brandon, S. Gowan, F. Boxall, W. Aherne, M. Rowlands, A. Hayes, V. Martins, F. Urban, K. Boxall, C. Prodromou, L. Pearl, K. James, T. P. Matthews, K. M. Cheung, A. Kalusa, K. Jones, E. McDonald, X. Barril, P. A. Brough, J. E. Cansfield, B. Dymock, M. J. Drysdale, H. Finch, R. Howes, R. E. Hubbard, A. Surgenor, P. Webb, M. Wood, L. Wright, P. Workman, *Cancer Res.* **2008**, *68*, 2850-2860.

[23] P. A. Brough, W. Aherne, X. Barril, J. Borgognoni, K. Boxall, J. E. Cansfield, K. M. Cheung, I. Collins, N. G. Davies, M. J. Drysdale, B. Dymock, S. A. Eccles, H. Finch, A. Fink, A. Hayes, R. Howes, R. E. Hubbard, K. James, A. M. Jordan, A. Lockie, V. Martins, A. Massey, T. P. Matthews, E. McDonald, C. J. Northfield, L. H. Pearl, C. Prodromou, S. Ray, F. I. Raynaud, S. D. Roughley, S. Y. Sharp, A. Surgenor, D. L. Walmsley, P. Webb, M. Wood, P. Workman, L. Wright, *J. Med. Chem.* **2008**, *51*, 196-218.

[32] Wright, L., Sheridan, L., Barril, X., Surgenor, A., Drysdale, M., Dymock, B., Collier, A., Massey, A., Beswick, M., Hubbard R.E. *Chem. Biol.* **2004**, *11*, 775-785.

REFERENCES

[1] G. J. Kleywegt, M. R. Harris, J. Y. Zou, T. C. Taylor, A. Wahlby, T. A. Jones, *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 2240-2249.

[2] J. J. Gasteiger, T. Engel, *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim, **2003**.

[3] B. W. Dymock, X. Barril, P. A. Brough, J. E. Cansfield, A. Massey, E. McDonald, R. E. Hubbard, A. Surgenor, S. D. Roughley, P. Webb, P. Workman, L. Wright, M. J. Drysdale, *J. Med. Chem.* **2005**, *48*, 4212-4215.

[4] L. Wright, X. Barril, B. Dymock, L. Sheridan, A. Surgenor, M. Beswick, M. Drysdale, A. Collier, A. Massey, N. Davies, A. Fink, C. Fromont, W. Aherne, K. Boxall, S. Sharp, P. Workman, R. E. Hubbard, *Chem. Biol.* **2004**, *11*, 775-785.

[5] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, *J. Phys. Chem.* **1993**, *97*, 10269-10280.

[6] D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman, M. Crowley, R. C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D. H. Mathews, C. Schafmeister, W. S. Ross, P. A. Kollman, *University of California*.

[7] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926.

[8] C. Jarzynski, *Phys. Rev. E. Stat. Nonlin Soft Matter Phys.* **2006**, *73*, 046105.

[9] L. Martinez, M. T. Sonoda, P. Webb, J. D. Baxter, M. S. Skaf, I. Polikarpov, *Biophys. J.* **2005**, *89*, 2011-2023.

[10] A. Fernandez, H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 113-118.

[11] A. Fernandez, A. Crespo, *Chem. Soc. Rev.* **2008**, *37*, 2373-2382.

3.3 Pocket prediction on proteins in motion

3.3.1 Introduction

"Lively proteins move and shake." The terminology of protein dynamics encompasses very different time-scales of molecular motion. From atomic vibration to protein folding, each level of motion has functional inference on the protein. Table 3.5 resumes characteristic motions, their corresponding time-scales and amplitudes.
Understanding the dynamics of a protein can give important insights into its

Time scale	Amplitude	Description
$1fs - 1ps$	0.001 - 0.1 $\AA$	bond stretching, angle bending constraint dihedral motion
$1ps - 1ns$	0.1 - 10 $\AA$	unhindered surface side chain motion loop motion, collective motion
$1ns - 1\mu s$	1 - 100 $\AA$	folding in small peptides helix coil transition
$1\mu - 1s$	10-100 $\AA$	protein folding

Table 3.5: Time scales of molecular motion, from `http://www.whatislife.com/reader/motion/motion.html`

mechanistic functioning. From this knowledge subsequent strategies can be derived aiming to modulate its function therapeutically. For example, allosteric effectors can be used to alter the binding affinity for the natural substrate of a protein [Rizk et al., 2011]. As another example, enzyme catalysis is known to depend on coupled thermal motions on the protein [Hammes-Schiffer and Benkovic, 2006].
A recent review on molecular mechanisms of action for discovered drugs highlights that several drugs alter protein motion [Swinney and Anthony, 2011]. Despite such evidence, few attempts are made in pharmaceutical industry to account for protein motion in rational drug-design. While the reasons for this reluctance won't be analysed here, they might be related to the complexity of protein motion.
Till today, no experimental or theoretical approach exists to gain a global insight into the conformational landscape of a protein. Thus, one is reduced to work with fragments, punctual snapshots in time of this landscape.

Computational methods exist to generate such fragments of the conformational landscape of a protein at atomistic resolution. Among them, molecular dynamics (MD). Physical principles of MD are explained in section 2.2. Here a method is presented that uses conformational ensembles to detect and characterise transient cavities and channels.

Pocket detection on moving proteins Identifying pockets on moving proteins is not a trivial task and surprisingly a flexible representation of a putative

pocket is still barely considered in today's drug discovery projects in pharmaceutical industry. As previously stated, a large panel of methods exist to predict pockets on static structures. Also, modern pocket prediction methods are very well performing and accurate to allow automated cavity prediction. Despite these advances, only one single method in the field considered the detection of transient cavities suitable for drug design. This method called EPOS^{BP} was introduced in 2007 by Eyrisch & Helms and is basically composed of running a sequence of cavity detections using PASS [Brady and Stouten, 2000] on conformations of a protein derived from a molecular dynamics trajectory. Till today, it remains the only method tackling this task, while new methods see the day allowing pocket prediction on static protein structures.
Nevertheless, the global approach chosen by Eyrisch and Helms is very valuable and was constituted the basics for the study presented in this section of the thesis.

Functional studies on moving proteins Interestingly, in a field not closely related to drug discovery, the consideration of protein motion and transientness of appearance and disappearance of cavities and channels is primordial to gain mechanistic insights into the functioning of various proteins. Such proteins usually contain internal channel systems transporting gas particles, but can also be transmembrane proteins transporting for example ions. In order to know if such a system could be of potential interest as pharmaceutical target, its mechanism of action has to be understood to be able to alter it. At this point it should be highlighted that internal channels as found in heme proteins, are very distinct from cavities targeted by structure based drug discovery.
In the past 10 years the study of channel motion and transientness has attracted substantially more contributors from academia, than the counter part on druggable binding sites. In order to study channels on dynamic and static structures, several computational tools exist. A description of the state of the art of those is provided in section 2.4.2 of this thesis.

3.3.2 Objectives

Development of a general purpose pocket tracking method for conformational ensembles The main objective of this work is to develop a software capable of transforming transient pocket and channel opening and closing observed on conformational ensembles of macromolecules to a visually easy to interpret representation. This transformation should be fast and relatively easy to produce.

Extraction of pocket characteristics on conformational ensembles Another central point of transient and mobile pocket detection is to know which protein conformer shows certain characteristics of interest. The program developed within this work should allow to track a given set of properties for a user defined zone on the protein surface. These properties could then further be used to either *(i)* characterize the cavity, *(ii)* to extract similar conformers having similar properties or *(iii)* to extract distinct conformers.

Identify conserved cavities on homologous proteins The method should be general and robust enough to allow for pocket identification on different homologous protein structures. The program should also be robust towards different protein lengths, missing atoms or residues.

3.3.3 Results

The pocket tracking method

Here a new method called MDpocket is presented allowing to track and characterise cavities and channels on conformational ensembles of proteins. As this work describes the development of a new method, the results section contains a description of the methods itself. MDpocket is based on the fpocket project described in section 2.3.5.
In order to analyse a conformational ensemble all conformers have to be structurally superimposed. Next, fpocket is run on all these conformations. While performing cavity detection a 3D grid is placed over the protein and each grid point tracks the *(i)* number of nearby alpha spheres and *(ii)* if at least one alpha sphere centre (Voronoi vertice) was nearby on each protein conformation. This workflow is schematically shown on figure 3.16. Once all pockets have been

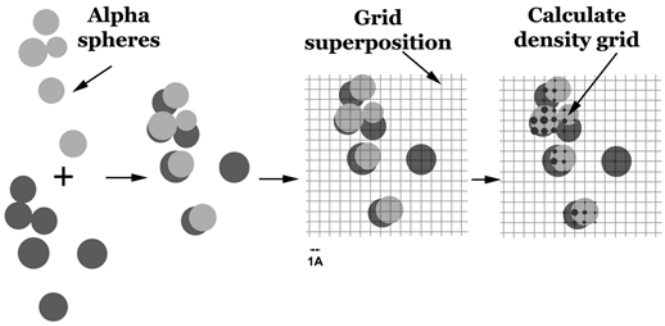


Figure 3.16: MDpocket workflow

assigned to the grid for all individual conformers, this grid can then be used to *(i)* visualise transient cavities and *(ii)* extract pocket/channel descriptors of a user-defined pocket. MDpocket produces 2 types of grids, *(i)* a frequency grid and *(ii)* a density grid.

Frequency grid This particular grid intends to capture how often a cavity or channel opened in a given region on the protein surface. It is normalized by the number of structures analysed. If transient pocket and channel openings occur within the time range of the conformational ensemble that is analysed, then this will be measurable in this grid as frequency of opening Φ_i

$$\Phi_i = \frac{1}{n} \sum_{i=1}^{n} \delta_i \tag{3.4}$$

where n corresponds to the number of snapshots analysed and δ_i to 1 if at least one alpha sphere centre was near to grid point i, 0 otherwise.

Density grid Protein flexibility and accounting for it in molecular simulations and analysis is very complex. Using the discrete way of detecting transientness of opening of pockets and channels in this method, we have to assume that within the conformational ensemble the cavity opens at least a certain amount of time. One way among others to produce conformational ensembles of protein structures is via molecular dynamics simulations. These simulations are often very time consuming to produce. In the end they allow for assessment of the conformational space of a protein corresponding to a nano second to micro second time scale. It is obvious that fortuitous openings of channels and cavities happening within that time interval could potentially be captured within the frequency map. However, very rare openings would not be included. In order to identify putative transient channels another type of grid is produced by MDpocket. This so-called density grid tracks the number of times an alpha sphere centre falls next to a given voxel on each conformer. The more atoms are packed, the higher the probability is to find more dense alpha sphere clusters. Higher atom packing is more frequently observed in concave and narrow environments like the lumen and the bottom of binding sites and within channels. Thus tracking the density of alpha spheres emphasizes also more buried zones of a channel or pocket.
The density map is derived using formula 3.5

$$\rho_i = \frac{1}{n} \sum_{i=1}^{n} d_{AS,i} \tag{3.5}$$

$$d_{AS,i} = card\left(AS_i \cap \{(x,y,z)_i \pm 0.5\}\right) \tag{3.6}$$

Visualising and interpreting pocket grids

As an example of how to visualise and interpret pocket grids produced by MDpocket, we selected 88 crystal structures of the heat shock protein 90 (HSP90). This protein was chosen because the known plasticity of the ATP binding site can be observed experimentally.

These known binding site movements are shown on figure 3.17 A (black arrow). The helix 4 - loop 2 - helix 5 motive can in some case form a straight helix opening a hydrophobic in the nearby ATP binding site. As the number of crystal structures used for this example is limited it is possible to count the number of structures being in conformation 1 (hydrophobic sub-pocket open) or conformation 2 (hydrophobic sub-pocket closed). In this example, 35.2% of all structures are of conformation 1, the rest of conformation 2.
All structures have been superimposed onto the reference structure (PDB code 1byq) and MDpocket was run producing *(i)* a frequency grid Φ and a density grid d. Results from MDpocket can be conveniently visualised as pocket frequency and density maps the same way as an electron density map. Popular visualisation software like VMD [Humphrey et al., 1996], Chimera [Pettersen et al., 2004] and PyMOL [DeLano, 2002] can be used. Figure 3.17 B shows

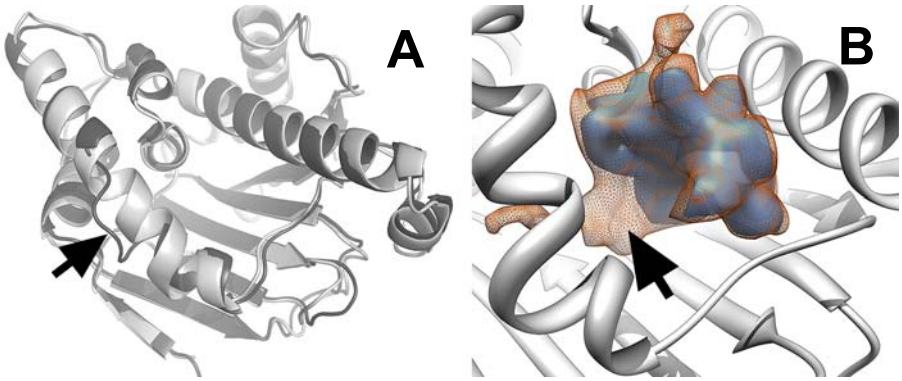


Figure 3.17: **A:** HSP90 helix motion lining the binding site, **B:** Frequency grids obtained with MDpocket on 88 crystal structures. Orange mesh: frequency 30% or more, blue surface: 40% or more pocket opening.

Φ at two levels of pocket opening, 30% (orange mesh) and 40% (blue surface). While the pocket frequency map evolves slowly on higher percentages, between 30 and 40% (more precisely around 35% a sub-pocket appearance can be observed. This subpocket, indicated by the black arrow on figure 3.17 B corresponds to the hydrophobic sub-pocket open in 35.2% of all input structures. This result is in excellent agreement with prior knowledge and shows how MDpocket frequency maps can be used on conformational ensembles of crystal structures or from MD trajectories.
The hydrophobic sub-pocket considered here can be of interest for drug design, as known inhibitors of HSP90 activity bind to this sub-pocket (figure not shown, but available as figure 1 related paper of MDpocket).
While frequency maps allow highlighting portions of binding sites that open frequently, the scale of simulation time and conformational space that is sampled with techniques like MD might be limited to observe very rare openings of transient cavities. Thus, such rare events are less likely to be easily visible within frequency maps. To circumvent this problem, another type of map is produced by MDpocket, tracking not the frequency, but the density of alpha spheres around a grid point during time. Transient openings tend to occur in closed environments characterised by high atom packing. Given the use of alpha spheres by the fpocket cavity prediction algorithm, the number of alpha spheres clustered in a given location are related to atom packing. Thus, such a density of alpha spheres can give insights into where transient and buried cavities open. In the example given in figure 3.17 MDpocket produced density maps allow to spot both the main binding site and the sub-pocket, although the sub-pocket is present only in 35.2% of all conformations.
MDpocket provides a overall structural cavity detection result, not precisely focusing on any zone. Via visual analysis, pocket grids allow thus a first exploratory view on overall transient pocket and channel openings. This is particularly helpful for identification not only of overall binding site motions, but also for discovery of unknown water access channels to the pocket.

Last, this exploratory cavity detection does not allow direct tracking or filtering of pocket properties. Thus, results provided can be sometimes difficult to analyse visually when lots of channels and cavities are found (typically heme proteins). Based on the work presented on druggability in this thesis, MDpocket however has the possibility to track cavities using the druggability score instead of alpha sphere presence and absence. In this mode, MDpocket is run and cavities are detected, but instead of mapping a pocket frequency to a grid, the druggability score is mapped to a grid. This allows tracking specific transient druggable pockets and might provide very useful for exploratory cavity detection on MD trajectories for identified putative allosteric binding sites.

Tracking a specific zone Despite the tracking for druggable pockets, the extraction of pocket properties is limited in the first step of the method. Thus, the user is allowed to select a zone of a density or frequency grid (with PyMOL for example) to delimit an area of interest. Once this area selected, it can be used as input for a subsequent MDpocket analysis, where all cavities are tracked, but only the cavity of interest selected by the user is analysed and various fpocket descriptors are extracted for each time-frame of the trajectory. Figure 3.18 shows an example of tracking the pocket volume of the ATP

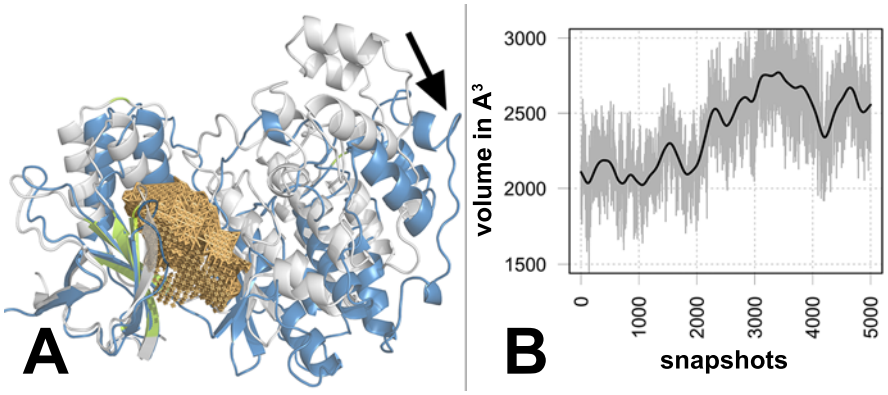


Figure 3.18: Tracking pocket properties with MDpocket. **A:** P38 Map kinase ATP binding site selected (orange grid). Two snapshots of a MD trajectory are shown, illustrating large lobe movement (black arrow). **B:** Volume of ATP binding site of P38 during MD trajectory calculated with MDpocket.

binding site of P38 Map kinase during a 50ns molecular dynamics trajectory. Here MDpocket was first run to produce a density grid. From the density grid the grid points represented in orange in figure 3.18 A have been selected with PyMOL. These grid points have then be used to extract properties of a zone the zone of interest. As an example, the volume of the cavity is shown in figure 3.18 B. Here we can observe that the average volume (thick black curve) is increasing especially after 2000 snapshots. Part A of the figure shows indeed that the big lobe of P38 is moving versus the small lobe (black arrow), knowing that the structural alignment has been done using the beta sheet lining the

binding site (green cartoon) as reference.

MDpocket use cases

Gas migration channel predictions As stated in the introduction, several methods have been proposed for detecting protein internal channels. MDpocket allows via it's versatility to detect small molecule binding sites but also smaller cavities and channels allowing diatomic ligand migration.
A well known example in the field of gaseous ligand migration is myoglobin. Small gaseous ligands like CO_2, O_2 or NO are known to bind to the heme of myoglobin. The system is very well studied especially with respect to potential pathways, gaseous ligands could use to enter the protein [Tilton et al., 1984, Scott and Gibson, 1997, Ostermann et al., 2000, Scott et al., 2001, Schotte et al., 2003, Cohen et al., 2006, Tomita et al., 2009]. Migration is assumed to be enabled via stable internal cavities connected by transient channels. To show the usefulness of MDpocket in the discovery of such transient channels, as well as stable cavities 10000 snapshots, equally spaced in time from a 50ns trajectory of myoglobin have been analysed. On figure 3.19 A results obtained

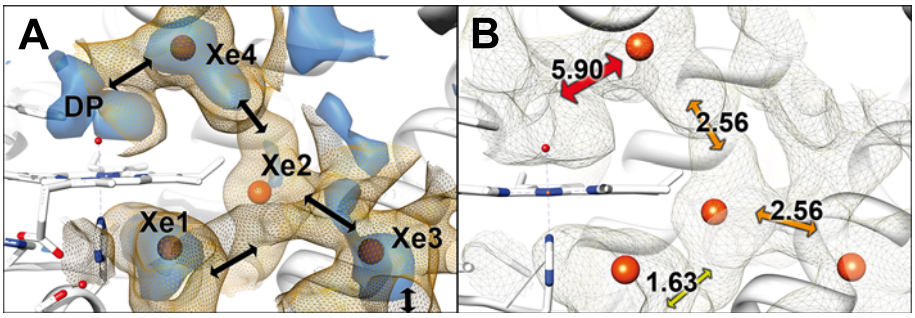


Figure 3.19: Myoglobin transient cavities detected with MDpocket from 50ns MD trajectory. **A:** Blue iso-surface, pocket frequency map at 60%m and orange mesh pocket density map at 2.0. **B:** Transient channel density. Known xenon (orange spheres) binding sites are found, as well as channels connecting them.

are shown and here again, different applications for frequency and density maps can be observed. While the frequency map at 60% pocket opening (blue iso-surface) shows known xenon binding sites, as well as the distal pocket above the heme and two other binding sites often found, the connections between them occur less frequently, such as they don't appear at this iso-level in the frequency map. To visualise them, the pocket density can be considered, as shown using the orange mesh in figure 3.19. Here clear connections can be observed between all stable binding sites, clearly indicating that these connect intermittently via these channels.
How and especially in which exact pathway these ligands migrate from one cavity to the other is still under scientific debate. However, it appears logical that wider and longer open transient channels would favour a faster exchange between sub-pockets and closed channels would hinder fast migration. To some

extent MDpocket can be used to favour certain routes versus others. Figure 3.19 B shows the minimum densities in alpha spheres when transient channels appear between stable cavities. Here, the connection between the distal pocket (DP) and xenon pocket Xe4 is found to be the most stable in agreement with time resolved X-ray crystallography experiments showing that the migration from DP to Xe4 is an initiating event [Schotte et al., 2003, Tomita et al., 2009]. Subsequent migration from Xe4 to Xe2 and Xe3 are observed. Here we find that Xe1 is predicted to be a stable pocket, connected by rarely opening transient channels to the outside and especially Xe2. Also in agreement with previously cited studies, showing high residence times of CO in Xe1, this results shows that MDpocket can provide qualitative and probably quantitative insights into protein internal ligand migration pathway and pathway networks.

Putative applications of MDpocket for ensemble docking The second use case presented for MDpocket is related to the study of binding sites of interest for drug discovery. Among usual steps in structure based drug discovery molecular docking is used to fit small organic molecules into the binding site of interest. This process called docking can be used to sieve through large databases of ligands and retrieve putative actives. While the ligand molecules usually are considered to be flexible during docking, the receptor structure is not. Several approaches exist to consider the complexity of including protein motion into small molecule docking, among which ensemble docking [Novoa et al., 2010]. Basically, several conformations of the protein are considered for docking all ligands. Given that docking is timely, using thousands of receptor conformations is prohibitive and an educated guess of which conformations to choose has to be made [Rueda et al., 2010].
Here MDpocket is used to analyse the previously used P38 ATP binding site. 32 crystal structures from known P38 binders co-crystallized with P38 have been systematically superimposed to 5000 snapshots from a 50ns MD trajectory of the same protein. This superimposition has been done using the stable beta-sheet lining the ATP binding site as reference. Ones superimposed, the ligand is extracted from the crystal structure and inserted into the MD snapshot and the interaction energy between the ligand and the protein is calculated using MOE. This process was done to simulate a docking process without all flaws in binding pose prediction and scoring that usually are paired with docking. These superimposed binding poses, by definition near to the cognate binding pose are then used to assess whether they fit correctly in the pocket or not. This is done considering that if steric clashes occur, the interaction energy is unfavourable (so positive). Thus, the amount of binders that have favourable interaction energies among the 32 binders can be calculated and tracked throughout the time of the MD simulation. This fraction of acceptable poses inside the moving pocket is represented as grey line curve on figure 3.20 and averaged as black curve.

As stated previously, several pocket descriptors can be extracted for an area of interest. Here, all MDpocket descriptors have been extracted and tested for correlation with the fraction of acceptable binding poses. Intuitively one might expect the pocket volume to correlate, relating more open and accessible

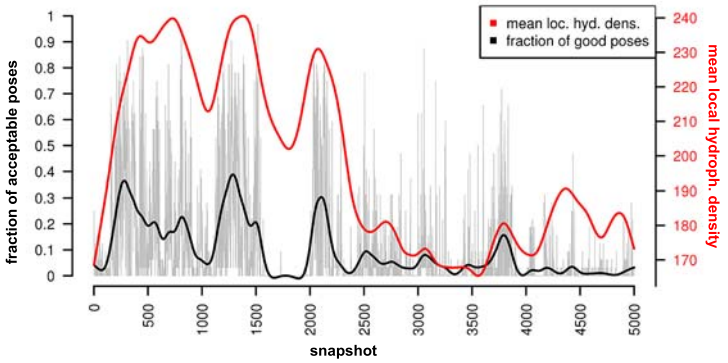


Figure 3.20: Fraction of binding poses without steric clashes (grey) from 32 known binders of P38 superimposed to 5000 snapshots of a MD trajectory. Average fraction is represented as thick black curve. The mean local hydrophobic density of the ATP binding site is shown in red.

binding sites to sterically acceptable binding poses of superimposed ligands. However, as could be seen in the analysis presented in figure 3.18, despite an important increase of the pocket volume after 2000 snapshots, the contrary of an improvement in binding poses can be observed. However, another descriptor, the mean local hydrophobic density is shown to correlate with the fraction of poses without steric clashes (red curve in figure 3.20). This descriptor has also been shown to be of importance in predicting druggability of a cavities in work presented previously in this thesis.
The mean local hydrophobic density allows to measure situations in time when hydrophobic sub-pockets and hydrophobic atom packing form dense local clusters in a binding site. While MD simulations in TIP3P water might hinder the exposure of large hydrophobic surface patches into the pocket, interaction between these and the ligand are important, especially seen the rather hydrophobic characteristics of drug-like molecules [Vieth et al., 2004]. Thus, tracking this descriptor allows to capture moments during the simulation when these hydrophobic surfaces are more accessible.

Novelty

MDpocket is the first free open-source tool allowing easy extraction of both, drug binding sites and transient channels from conformational ensembles. MDpocket is invariant to the source of these conformations and can thus be used to analyse MD trajectories or homologous protein structures. Compared to EPOS^{BP}, MDpocket provides a continuous description of cavities and does not split them according to clustering rules. This is of importance when studying the whole spectrum of descriptors of a cavity of interest.

Limitations

Despite advantages of MDpocket, several limitations exist. The most important drawback is due to the methodology and the necessary superimposition. Such a superimposition is by definition erroneous on conformational ensembles. Thus, structural alignments have to be done with a focus on parts of the structure that are either stable or of particular interest.
Another limitation is linked to the relative length of MD simulations. In order to be able to observe and measure the frequency of opening of transient channels in the frequency map a substantial simulation length and number of snapshots might be required. Such simulations lengths are currently barely accessible.
Last, regarding channel predictions, MDpocket does not allow measurements of channel dimensions.

Bibliography

G P Brady and P F Stouten. Fast prediction and visualization of protein binding pockets with PASS. *Journal of computer-aided molecular design*, 14(4):383–401, May 2000.

Jordi Cohen, Anton Arkhipov, Rosemary Braun, and Klaus Schulten. Imaging the migration pathways for O2, CO, NO, and Xe inside myoglobin. *Biophysical journal*, 91(5):1844–57, September 2006.

Warren L DeLano. The PyMOL Molecular Graphics System, 2002.

Sharon Hammes-Schiffer and Stephen J Benkovic. Relating protein motion to catalysis. *Annual review of biochemistry*, 75:519–41, January 2006.

W Humphrey, A Dalke, and K Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–8, 27–8, February 1996.

Eva Maria Novoa, Lluis Ribas de Pouplana, Xavier Barril, and Modesto Orozco. Ensemble Docking from Homology Models. *Journal of Chemical Theory and Computation*, 6(8):2547–2557, August 2010.

A Ostermann, R Waschipky, F G Parak, and G U Nienhaus. Ligand binding and conformational motions in myoglobin. *Nature*, 404(6774):205–8, March 2000.

Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–12, October 2004.

Shahir S Rizk, Marcin Paduch, John H Heithaus, Erica M Duguid, Andrew Sandstrom, and Anthony A Kossiakoff. Allosteric control of ligand-binding affinity using engineered conformation-specific effector proteins. *Nature Structural Molecular Biology*, 18(4):437–442, 2011.

Manuel Rueda, Giovanni Bottegoni, and Ruben Abagyan. Recipes for the selection of experimental protein conformations for virtual screening. *Journal of Chemical Information and Modeling*, 50(1):186–193, 2010.

Friedrich Schotte, Manho Lim, Timothy A Jackson, Aleksandr V Smirnov, Jayashree Soman, John S Olson, George N Phillips, Michael Wulff, and Philip A Anfinrud. Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science (New York, N.Y.)*, 300(5627):1944–7, June 2003.

E E Scott and Q H Gibson. Ligand migration in sperm whale myoglobin. *Biochemistry*, 36(39):11909–17, September 1997.

Emily E Scott, Quentin H Gibson, and John S Olson. Mapping the Pathways for O2 Entry Into and Exit from Myoglobin. *Journal of Biological Chemistry*, 276(7): 5177–5188, 2001.

David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews. Drug discovery*, 10(7):507–19, January 2011.

Robert F. Tilton, Irwin D. Kuntz, and Gregory A. Petsko. Cavities in proteins: structure of a metmyoglobin xenon complex solved to 1.9 .ANG. *Biochemistry*, 23 (13):2849–2857, June 1984.

Ayana Tomita, Tokushi Sato, Kouhei Ichiyanagi, Shunsuke Nozawa, Hirohiko Ichikawa, Matthieu Chollet, Fumihiro Kawai, Sam-Yong Park, Takayuki Tsuduki, Takahisa Yamato, Shin-Ya Koshihara, and Shin-Ichi Adachi. Visualizing breathing motion of internal cavities in concert with ligand migration in myoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8): 2612–6, February 2009.

Michal Vieth, Miles G Siegel, Richard E Higgs, Ian A Watson, Daniel H Robertson, Kenneth A Savin, Gregory L Durst, and Philip A Hipskind. Characteristic physical properties and structural fragments of marketed oral drugs. *Journal of medicinal chemistry*, 47(1):224–32, January 2004.

MDpocket : Herramienta de Detección de Cavidades y Caracterización de Trayectorias en Dinámica Molecular

Peter Schmidtke, Axel Bidon-Chanal, F. Javier Luque y Xavier Barril
en revision, Bioinformatics

Motivación: Una variedad de algoritmos para la detección de cavidades son ahora gratuitos o están comercialmente disponibles para la comunidad científica, para el análisis de estructuras de proteínas estáticas. Sin embargo, desde que las proteínas son entidades dinámicas, aumentan las capacidades de dichos programas para, de manera sencilla, detectar y caracterizar las cavidades teniendo en cuenta los conjuntos conformacionales valuables para capturar la plasticidad de las cavidades, y por tanto, permiten mejorar la visión de las relaciones estructura-función.
Resultados: Este trabajo describe un nuevo método, llamdo MDpocket, que proporciona una nueva herramienta rápida, gratuita y de código abierto, para el seguimiento de moléculas pequeñas, sitios de unión y migración de gas en las trayectorias de MD u otras agrupaciones conformacionales. MDpocket está basado en el algoritmo de detección de cavidades, además de una valorable contribución de herramientas de análisis pre-existentes. Las capacidades de MDpocket están ilustradas desde tres casos relevantes: i) la detección de subhuecos transitorios utilizando un complejo cristalizado de estructuras HSP90 ii) la detección de conocidos sitios de unión de xenon y vías de migración en la mioglobina y iii) la identificación de huecos por Docking molecular en Map quinasa P38.
Disponibilidad: MDpocket es gratuito y el software de código abierto puede ser descargado desde: `http://fpocket.sourceforge.net`.

STRUCTURAL BIOINFORMATICS

MDpocket : Open Source Cavity Detection and Characterization on Molecular Dynamics Trajectories

Peter Schmidtke[1,*], Axel Bidon-Chanal[2], F. Javier Luque[1] and Xavier Barril[1,3]

[1] Departament de Fisicoquímica and Institut de Biomedicina (IBUB), Facultat de Farmàcia, Universitat de Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain. [2] Equipe de Dynamique des Assemblages Membranaires, UMR No. 7565, Centre National de la Recherche Scientifique-Université Henri Poincaré, Nancy, France, [3] Institució Catalana de Recerca i Estudis Avanats (ICREA)

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: A variety of pocket detection algorithms are now freely or commercially available to the scientific community for the analysis of static protein structures. However, since proteins are dynamic entities, enhancing the capabilities of these programs for the straightforward detection and characterization of cavities taking into account protein conformational ensembles should be valuable for capturing the plasticity of pockets, and therefore allow gaining insight into structure-function relationships.

Results: This paper describes a new method, called MDpocket, providing a fast, free and open source tool for tracking small molecule binding sites and gas migration pathways on MD trajectories or other conformational ensembles. MDpocket is based on the fpocket cavity detection algorithm and a valuable contribution to existing analysis tools. The capabilities of MDpocket are illustrated for three relevant cases: i) the detection of transient sub-pockets using an ensemble of crystal structures of HSP90 ii) the detection of known xenon binding sites and migration pathways in myoglobin, and iii) the identification of suitable pockets for molecular docking in P38 Map kinase.

Availability: MDpocket is free and open source software and can be downloaded at http://fpocket.sourceforge.net.

Contact: pschmidtke@ub.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Over the past two decades a variety of algorithms have been proposed with the aim to identify binding pockets for small molecules in biomolecular targets (An, Totrov, & Abagyan, 2005; Brady & Stouten, 2000; Hendlich, F. Rippmann, & Barnickel, 1997; D. Kim et al., 2008; Kleywegt & Jones, 1994; Laskowski, 1995; Laurie & R. M. Jackson, 2005; Le Guilloux, Schmidtke, & Tuffery, 2009; Peters, Fauck, & Frömmel, 1996; Weisel, Proschak, & G. Schneider, 2007). These algorithms can be classified in three broad classes depending on the general method used for cavity detection: (i) geometry-based, (ii) energy-based or (iii) sequence-based methods. The first class relies on geometrical features of pockets with no or few considerations for the interaction energy between a putative ligand and the pocket. Energy-based algorithms estimate the suitability of a pocket to bind a molecule using probe-pocket interaction energy calculations. The latter methods are usually computationally more expensive and require specific atom typing and the use of underlying force fields. Sequence-based methods exploit the propensity of conserved residues in the binding site. Last, various hybrid methods, combining at least two of the previous approaches, also exist (Halgren, 2007; B. Huang, 2009; Liang, Edelsbrunner, & Woodward, 1998). The reader is directed to recent reviews for a detailed discussion of different cavity detection algorithms (Henrich et al., 2010; Leis, S. Schneider, & Zacharias, 2010; Pérot, Sperandio, Miteva, Camproux, & Villoutreix, 2010).

The vast majority of cavity detection algorithms have been developed to treat static structures, like crystal structures of proteins available in the PDB. However, this represents a serious limitation to account for the intrinsic plasticity of the binding pocket. Protein dynamics act on a multitude of aspects in protein function. For instance, side chain flipping or domain motions can obstruct or free internal cavities or channels that allow migration of ligands, and reshape the binding sites (Axel Bidon-Chanal et al., 2006; Carrillo & Orozco, 2008; Spyrakis et al., 2011). In turn, these findings raise challenging questions about the impact of protein flexibility on the topological features of cavities and their binding properties.

Few works have attempted to account for the dynamical behaviour of proteins in the identification of binding cavities and tunnels. The interplay between protein dynamics and ligand migration pathways can be characterised by tools that rely on prior molecular dynamics (MD) simulations and further post-processing of the trajectory. VOIDOO (Kleywegt & Jones, 1994) allows internal cavity and volume calculations, but it is rather time consuming and its use is not straightforward. CAVER (Beneš et al., 2010) is a PyMOL plugin that allows internal channel detection on MD trajectories. CAVER was recently improved to a software called MOLE using computational geometry principles instead of grid-based calculations (Petrek, Kosinová, Koca, & Otyepka, 2007). More recently Wolfson and coworkers have proposed a method called MolAxis for detection of channels from the interior of the protein to the bulk

*To whom correspondence should be addressed.

solvent (Yaffe, Fishelovitch, Wolfson, Halperin, & Nussinov, 2008). These methods are designed to detect channels on static structures or conformational ensembles of the protein.

In order to examine the suitability of internal pathways for ligand migration, more time consuming techniques like implicit ligand sampling have been proposed (Cohen, Olsen, & Klaus Schulten, 2008). Glazer et al. have used MD for the identification of calcium binding sites and have shown that the inclusion of dynamic behaviour can improve function prediction (Glazer, Radmer, & Altman, 2009). Recently a MATLAB-based approach called DyME (Lin & Song, 2011) has been proposed for analysis of putative internal channels from MD trajectories using Voronoi tessellation and clustering techniques. Regarding the binding of ligands in cavities, protein dynamics has been accounted for by using a set of privileged static structures, which are chosen as representative conformational states of the pocket based on experimental X-ray structures or from MD simulations of the target (Barril & Morley, 2005; Kua, Zhang, & McCammon, 2002; Novoa, Pouplana, Barril, & Orozco, 2010). More recently, Eyrisch and Helms established a protocol for the detection of transient cavities in protein-protein interfaces by using MD simulations and the cavity detection algorithm PASS (Brady & Stouten, 2000; Eyrisch & Helms, 2007). Other approaches combine conformational sampling and/or selection from MD simulations to allow pocket adaptation to a given ligand positioned in the binding site (Sherman, Day, Jacobson, Friesner, & Farid, 2006). Finally, a distinct strategy has been adopted in PELE (Borrelli, Vitalis, Alcantara, & Guallar, 2005), as internal pockets and channels are identified based on localized perturbations that combine Monte Carlo sampling and energy minimization calculations to track and predict ligand migration pathways.

In this work a new generic pocket detection program called MDpocket is presented. The aim of MDpocket is to identify and characterize binding sites and channels that might be transiently formed in the protein from the analysis of conformational ensembles generated by MD or other sources. The core of this new program is the recently published open source platform Fpocket (Le Guilloux, Schmidtke, & Tuffery, 2009), which is a very fast geometry-based cavity detection algorithm. The platform relies on three programs: i) *fpocket*, which identifies cavities in a protein, ii) *dpocket*, which extracts descriptors of the pocket, and finally iii) *tpocket*, which allows assessment of pocket scoring functions. One main advantage of *fpocket* is its adaptability to a given problem: initially developed for the discovery of small molecule binding sites, it can detect different types of cavities, including very small pockets, ligand binding sites or even tunnels via a proper choice of parameters available through command line.

MDpocket is fast and well suited for the study of processes where tracking of cavities is of interest. In particular, the capabilities of MDpocket are illustrated by examining i) the plasticity of the HSP90 binding site using a set of X-ray crystallographic structures, ii) the analysis of xenon binding sites and migration pathways in myoglobin, and iii) the selection of suitable docking sites for P38 DFG-in binders.

The software is freely available as part of the Fpocket software package and can be downloaded from http://fpocket.sourceforge.net.

2 METHODS

2.1 MDpocket input

The general input format of MDpocket is a text file listing filenames to all pdb files to be considered for the analysis. This choice is motivated by the fact that MD trajectories are stored in different file formats depending on the specifications defined in programs such as Amber (Case et al., 2005), Charmm (Brooks et al., 2009; MacKerel Jr. et al., 1998), Gromacs (Hess, Kutzner, van der Spoel, & Lindahl, 2008) or NAMD (J. C. Phillips et al., 2005). Due to the lack of a common format, we have decided to transform the trajectory to a set of pdb files corresponding to snapshots taken along the simulation. Moreover, when those files are ordered by time, MDpocket permits the analysis of time-dependent events. In addition, the use of pdb files also facilitates the analysis of X-ray structures taken from the PDB. Finally, the PDB files do not need to be identical (no generic topology required), which makes MDpocket easily adaptable to analyse conformational ensembles from various sources (homologous proteins, for instance).

To carry out the cavity detection with MDpocket, it is important to superimpose the PDB structures onto each other. To this end, solvent molecules and counter ions were stripped off the system prior to pdb export, and then structural alignments were carried out using ptraj from AmberTools (Case et al., 2005).

2.2 *fpocket* parameters and output

MDpocket relies on the pocket detection program *fpocket*, which makes extensive use of Voronoi tesselation during cavity detection. This geometric approach allows retrieving the alpha spheres (i.e., spheres that are in contact with exactly 4 atoms without any other atom situated within the sphere). The center of the alpha sphere corresponds to a Voronoi vertex. A list of all Voronoi vertices (clustered into pockets) situated on the protein surface is provided in the output of *fpocket*.

The *fpocket* module is very flexible regarding the type of cavity to be detected. The flexibility is achieved through user accessible command line parameters that influence filtering and clustering of alpha spheres. The most important parameters are those that define the size of alpha spheres built up in a binding site (-m : minimum alpha sphere size, -M : maximum alpha sphere size). Moreover, filtering and clustering of alpha spheres can be modified using parameters –i (the minimum number of alpha spheres in the final pocket) and –n (the minimum number of alpha spheres close to each other for merging two binding sites into a single one).

Three different parameter sets for pocket detection have been assessed here in order to illustrate the scalability of the algorithm,. Set 1 denotes the default *fpocket* parameter set (-m 3.0, -M 6.0, -i 30, -n 3), which is tailored for detection of small molecule (i.e., peptides, drug-like compounds) binding sites. Set 2 is intended to identify very small channels and pockets (-m 2.8, -M 6.0, -i 3, -n 2). Finally, set 3 (-m 3.5, -M 5.5, -i 1, -n 2) is chosen to represent an alpha sphere with a physically meaningful minimum size, while retaining all the pockets (even tiny ones built by a single alpha sphere), it being thus better suited to identify very open cavities physically accessible to a water molecule and detection of continuous channels that can accommodate a water molecule.

2.3 MDpocket workflow: pocket detection

Pockets are detected based on the workflow depicted in Figure S1:

a) A 1Å spaced grid is placed over the first snapshot of the set of superposed pdb files.

b) *fpocket* is run on every snapshot of the set of pdb files.

c) For every snapshot each alpha sphere is assigned to the grid point *i* closest to the alpha sphere centre. Note that several alpha spheres can be assigned to a given point of the grid originating from the same snapshot or

from different snapshot. The number of counted alpha spheres assigned to each grid point is then normalized by the number of snapshots in order to generate the density map, ρ (Eq. 1).

$$\rho_i = \frac{1}{n}\sum_{i=1}^{n} d_{AS,i} \quad (1)$$

where n corresponds to the number of snapshots, AS stands for alpha sphere, and $d_{AS,i}$ is determined as

$$d_{AS,i} = card\left(AS_i \cap \left\{(x,y,z)_i \pm 0.5\right\}\right) \quad (2)$$

where the tolerance of 0.5 is considered for all grid points.

d) In addition, for every snapshot each grid point i is given an occupancy parameter δ_i, which equals 1 or 0 depending on the previous assignment of alpha spheres to that point (δ_i = 1 if at least one alpha sphere has been assigned to the grid point i, otherwise δ_i = 0). Then, a frequency map, ϕ, is generated by normalizing the sum of values δ_i assigned to point i by the total number of snapshots (Eq. 3).

$$\Phi_i = \frac{1}{n}\sum_{i=1}^{n} \delta_i \quad (3)$$

The normalized pocket frequency map (Eq. 3) allows visualisation of the opening frequency of a pocket during a MD trajectory. Thus, it indicates whether a given point in the grid is permanently accessible (Φ_i = 1), hindered (Φ_i = 0), or transiently accessible (0 < Φ_i <1). In contrast, the pocket density map (Eq. 1) is intended to provide information about the environmental atom packing around the pocket. Both density and frequency maps can be visualized using VMD (Humphrey, Dalke, & K Schulten, 1996), Chimera (Pettersen et al., 2004) or PyMOL from the output files produced by MDpocket.

In contrast to the pocket detection on a single static structure, MDpocket provides information about the plasticity of pockets from the normalized frequency/density maps generated from a given ensemble of structures. This is a major difference to other approaches that assign discrete pocket Ids to track pockets during MD trajectories (Eyrisch & Helms, 2007). In MDpocket an accurate pocket Id identification and tracking is not necessary, rendering the detection protocol more generic and less error prone.

2.4 MDpocket workflow: pocket characterization

Frequency and density maps are valuable to explore pocket opening/closure. MDpocket also permits to characterize those pockets or binding sites by providing a variety of descriptors, which include the accessible surface area and volume of the pocket, the number of alpha spheres, and the mean local hydrophobic density, which is an index of binding site druggability (Schmidtke & Barril, 2010).

To carry out the pocket characterization, the user can extract all grid points having a grid value equal or higher than a certain threshold from the previously calculated pocket frequency map (a default value of 0.5 is defined in MDpocket). Thus, visualization of the frequency map permits the user to select an area of interest (i.e., a transient channel or a binding site) using a graphical display tool, and the user-defined zone (saved as pdb file) can then be used as input for MDpocket in order to determine all pocket descriptors corresponding to the selected area for the whole ensemble.

2.5 MDpocket validation

The usefulness and accuracy of MDpocket have been calibrated considering three molecular systems studied previously in our group.

HSP90. A 78.5 ns trajectory run of the N-terminal domain of the heat shock protein 90 (HSP90) with explicit solvent (TIP3P water model) was first considered. The simulation was run in the NPT ensemble (1 atm, 298 K) using periodic boundary conditions and Ewald sums (grid spacing of 1 Å) for long-range electrostatic interactions. The parm99 force field and the Amber (Case et al., 2005) package were also used. From this trajectory 3925 equally spaced snapshots were extracted and analyzed with MDpocket. Furthermore, an alternative ensemble of structures was built up by retrieving 88 X-ray crystallographic structures from the PDB (see Table S1), which were subsequently aligned using PyMOL.

Myoglobin. The crystal structure of myoglobin (PDB entry 1VXD) (F. Yang & G N Phillips, 1996) was immersed in an octahedral box of TIP3P water molecules and the net charge of the system was neutralized with sodium ions. The final system contained around 21000 atoms. The simulations were run using the PMEMD module of amber9 and the parmm99 force field with special parameters for the haem residue (Bidon-Chanal et al., 2006; Marti et al., 2006). The SHAKE algorithm was used to keep bonds involving hydrogen atoms at their equilibrium length, in conjunction with a 1 fs time step for the integration of the Newton's equations. Trajectories were collected in the NPT ensemble (1 atm, 298 K) using periodic boundary conditions and Ewald sums (grid spacing of 1 Å) for long-range electrostatic interactions. The systems were minimized using a multistep protocol, involving first the adjustment of hydrogens, then the refinement of water molecules, and finally the minimization of the whole system. The equilibration was performed by heating from 100 to 298 K in four 100-ps steps at 150, 200, 250 and 298 K. Finally, a 50 ns trajectory was obtained, collecting frames at 1 ps intervals. The MDpocket analysis was performed with 10000 snapshots equally spaced in time.

P38 Map kinase. The PDB structure 1P38 was used as initial structure for a 50 ns MD trajectory. Leap was used to immerse the protein in an octahedral solvent box. The overall charge of the system was neutralized by addition of counterions. The solvent box contained a mixture of water and 20% isopropanol molecules. In order to obtain more information about the equilibration protocol, refer to Seco et al. (Seco, Luque, & Barril, 2009). The production run was carried out at 1atm and 300K using periodic boundary conditions. 5000 snapshots equally spaced in time have been used for the MDpocket analysis.

To assess whether MDpocket is able to give useful hints during the selection of receptor conformations for molecular docking, 32 X-ray crystallographic structures of P38 with DFG-in conformations and bound ligands were extracted from the PDB (see List S2), and aligned to all snapshots of the MD using the C$_\alpha$ atoms of residues 35-39, 45-50 and 100- 104, which correspond to the stable part of the beta sheet lining the binding site. In order to extract the interaction energies for each DFG-in ligand, the aligned ligand was extracted from the crystal structure and added to each snapshot of the MD trajectory to calculate the interaction energy. All energy calculations were performed using MOE (Chemical Computing Group, 2009), and the default potential energy function with the MMFF force field. No modifications or conformational changes were applied to the ligands near the residues in the binding site. Thus, this very crude interaction energy evaluation should mainly give insights into steric clashes that could occur in the ligand-protein complex, if the ligand is docked in a given conformation of the protein sampled during the MD trajectory.

3 RESULTS

3.1 Spotting and interpreting variation on structural ensembles

Although protein dynamics and flexibility can be crucial for recognition with other proteins or small ligands, it is still an often-neglected aspect in structural analysis. Insights into protein flexibility can be gained from the structural differences observed in X-ray structures or by using NMR-derived data. In both cases, expensive equipment is needed and sample preparation can be a real handicap. On the other hand, MD represents a powerful theoretical tool to explore the dynamical behaviour of biomolecules and MDpocket is intended to take advantage of this fact to identify and characterize pockets and binding sites in proteins from the analysis of structural ensembles chosen to account for the intrinsic flexibility of proteins.

The capability of MDpocket is first explored by considering the heat shock protein 90 (HSP90), which was chosen because the available X-ray data reveals a transient opening of a hydrophobic sub-pocket connected to the known ATP binding site. This opening occurs when loop 2 (residues N105 – A111) in the helix 4–loop 2–helix 5 motif reorganises to an alpha helix (Wright et al., 2004). 88 different PDB structures (see Table S1) were superimposed to the reference structure 1BYQ using PyMol. In this set 31 structures have a straight helix conformation (corresponding to the open sub-pocket), whereas the rest of the conformations show predominance of loop 2 and thus the sub-pocket does not exist. Next, MDpocket was run on all superimposed structures.

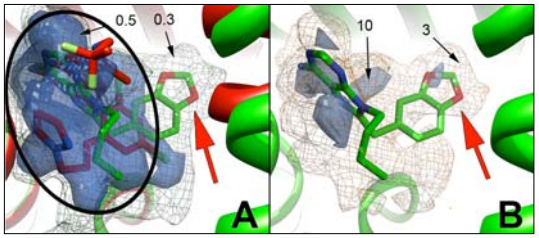


Fig. 1. HSP90 binding site derived from 88 X-ray crystallographic structures. (A) MDpocket pocket frequency map at 50% (blue iso-surface) and 30% (mesh). The red structure corresponds to a crystal structure where the sub-pocket is closed. The green structure has the sub-pocket open. The main pocket (black ellipse) is found in all snapshots. The subpocket (red arrow) is open in 35.2% of all X-ray structures, and the isosurface determined from MDpocket (mesh) appears at 35% of pocket appearance frequency. (B) MDpocket pocket density map (for clarity only the green structure is shown) at two levels of pocket density. The main pocket is found at low (3, yellow mesh) and high (10, blue surface) densities. The subpocket is also found at low (3) densities and a spot (blue surface) can also be seen at high densities despite the fact that the pocket does not open frequently.

Figure 1A shows the HSP90 binding site with a ligand in the main binding site (red protein and ligand) and another ligand filling the open sub-pocket (red arrow). The pocket frequency map, which gives the amount of time a pocket was found, is represented at two different isocontours (0.5, blue isosurface; 0.3, orange mesh). At the lower isovalue (0.3, corresponding to minimum 30% pocket opening on all conformations), the subpocket is open. The isovolume corresponding to the subpocket disappears above 35% and is not visible anymore for 50% of all conformations (blue isosurface). This result is in agreement with our prior knowledge that the pocket is open in about 35.2% of all conformations.

Figure 1B illustrates the pocket density map, which provides information on the density of alpha spheres in a pocket. The density map is represented at very high (10, blue isosurface) and low (3, orange mesh) isocontours. Interestingly, one can observe a highly dense isosurface in the subpocket (red arrow), although it just opens 35.2% in all conformations. Thus, the density map can give useful insights about the relative enclosure of different regions of a binding site regardless of its frequency of appearance.

3.2 Influence of the structural alignment on the results

The detection of pockets with MDpocket can be affected by the structural alignment of the frames included in the ensemble of structures. The influence of structural alignment on the results has been explored for 3925 snapshots taken from a 50ns trajectory of P38 Map kinase, which was chosen due to the flexibility of the two lobes that define the ATP binding site (see Figure 2). Thus, Figure 2 shows how the alignment of the small lobe in snapshots taken at the beginning and end of the trajectory leads to a large displacement in the bigger lobe, which reflects the relative motion between the two lobes along the trajectory.

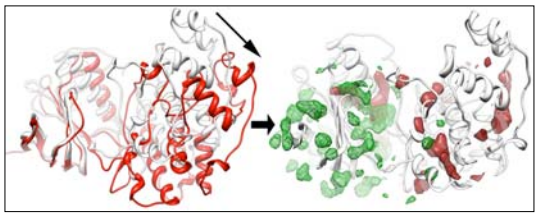


Fig. 2. Representative snapshots showing the potential effect of alignment on MDpocket results for P38 Map kinase. Left: An important motion of one lobe versus the other is observed during the MD trajectory (all snapshots have been aligned using the small lobe as reference), white ribbon - beginning of the trajectory, red ribbon - end of the trajectory. Right: Pocket frequency maps at 40% of pocket opening for (green) a trajectory aligned with residues lining the active site, and (red) a trajectory aligned with all residues.

Next, two scenarios have been considered. In the first case the superimposition of all structures was done using the Cα atoms of all protein residues (global alignment). Second, only the Cα atoms of a stable part of the small lobe (the beta sheet lining the binding site; local alignment) were considered for superimposition. The MDpocket frequency maps at an isocontour of 0.4 reveals major differences between the two alignments regarding the pocket appearance. The green mesh in Figure 2 displays MDpocket results derived for the local alignment. Several conserved pockets can be observed on the small lobe and also the active site is well identified. However, no clear pocket is found on the bigger lobe. This is due to the large inter-lobe motions occurring during the MD. The

red isosurface reflect the same isocontour derived for the global alignment. In this case, pockets can be found on the bigger lobe, but the active site is just partially identified at this level of pocket frequency.

This example illustrates the sensitivity of the grid-based methodology implemented in MDpocket, especially when large motions between lobes or domains are involved, as one should expect that these changes will have a marked influence on the pockets. As a rule-of-thumb, it can be stated that if the aim of the MDpocket analysis is the study of one particular cavity or channel, a superimposition using the heavy atoms lining the pocket should be used. Otherwise, usage of all heavy atoms on the protein can be considered though mobile parts should be excluded for the alignment, as, pockets in contact with these mobile parts can yield underestimated pocket frequencies in MDpocket analysis.

3.3 Effect of *fpocket* parameters on the results

The proper choice of cavity detection parameters can be decisive for obtaining an accurate description of both tiny channels transiently formed in the protein matrix or permanent drug-like binding sites within the protein. In order to tackle these distinct situations, three different parameter sets (see Methods) have been used to analyse the HSP90 in the region enclosing the binding site. Parameter set 1 (default) is intended to identify binding sites able to bind small molecular substrates or drug-like molecules. Results obtained for this parameter set are shown as green mesh in Figures 3A and 3B for the ensemble of 3925 HSP90 snapshots. Parameter set 2 is conceived for the identification of internal channels and small molecule binding sites (results shown as red isosurface in Figure 3A). Finally, parameter set 3 is intended to support detection of cavities sterically able to host diatomic ligands or water molecules (results shown in Figure 3B).

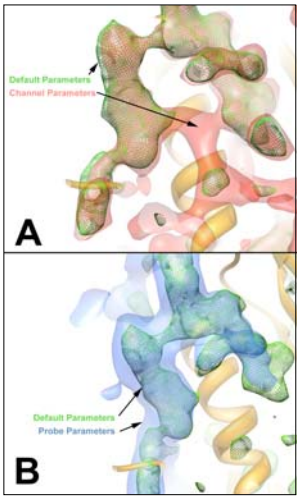


Fig. 3. MDpocket isosurfaces/meshes for (A) parameters set 1 (default parameters; isovalue 0.4) and set 2 (intended for channel parameters; isovalue 0.8), and (B) parameters set 1 (isovalue 0.4) and set 3 (water probe; isovalue 0.25)

Inspection of Figure 3 reveals differences in the shape of the cavities/tunnels delineated by the isocontours generated from the three parameter sets. Thus, Figure 3A clearly shows the suitability of parameter set 2 to identify internal channels, which are nevertheless found as discontinuous regions when parameter set 1 is used. Interestingly, the putative channel observed on figure 3A corresponds to a region that is known to be part of a larger loop, opening during the chaperoning cycle (Ali et al., 2006). Likewise, Figure 3B shows that parameter set 3 discloses the solvent exposed part of the binding site, while very tiny internal channels or narrow parts of internal pockets are not identified here.

Hence, it can be concluded that putative gas or water migration pathways can be identified using parameter set 2. Indeed, a putative channel is found below the sub-pocket discussed earlier.

Even though parameter set 2 allows identification of tiny cavities and channels, they can be however physically meaningless if no molecule can fill them. Thus, calibration against pockets and channels derived using a specific probe mimicking the molecule of interest (i.e., a diatomic molecule or a water) can be valuable to gain insight into ligand migration. However, if the main interest is to track transient druggable cavities on the protein surface, default parameters (set 1) are more adequate. In any case, the user can easily adjust the parameters required for pocket detection, thus facilitating the exploration towards specific class of binding sites and channels.

3.4 Case 1: Pocket detection in myoglobin

Myoglobin (Mb) is known to bind small diatomic ligands such as CO, O_2 or NO. Different internal cavities suited to hold small diatomic ligands were first detected in soaking experiments of Mb crystals with Xe atoms (Tilton, Kuntz, & Petsko, 1984). A large number of studies have been performed to characterise the ligand migration in Mb (Cohen, Arkhipov, Braun, & Klaus Schulten, 2006; Ostermann, Waschipky, Parak, & Nienhaus, 2000; Schotte et al., 2003; E E Scott & Q H Gibson, 1997; Emily E Scott, Quentin H Gibson, & John S Olson, 2001; Tomita et al., 2009). It is generally assumed that migration involves an intermittent passage between transient pockets, and the ligand entry to the distal cavity from the solvent phase. Here our aim is to show that MDpocket is a useful tool to detect those preferential Xe binding pockets, to identify the pathways that connect them in the interior of Mb, and eventually detect the possible entrance for the ligand from the solvent phase from an ensemble of snapshots taken from a MD simulation.

The parameter set 2 was used to analyse 10000 snapshots evenly taken over a 50 ns MD trajectory. As expected, the results show that frequently appearing pockets overlap with known Xe binding sites in the crystal structure (see blue isosurface in Figure 4), which indicates that the cavities are present in the majority of the snapshots. Thus, they should generally be also detectable when only a single snapshot is used for the analysis. On the other hand, connections between cavities do not occur through long-lived channels, i.e. they are not present in the ensemble of crystal structures deposited in the PDB, which is in agreement with the notion that migration of diatomic ligands is triggered by transient opening of channels between pockets (Ostermann, Waschipky, Parak, & Nienhaus, 2000; Schotte et al., 2003). Using conformations coming from the MD trajectory, one can observe sparse opening of these channels with MDpocket. As these are characterised by a short lifetime, they are identified as less frequently appearing pockets. As seen in the HSP90 example, MDpocket is capable of detecting these transient channels and visualizing them using pocket density maps. Thus, the orange isomesh shown in Figure 4 delineates putative migration pathways between pocket sites. A more precise display can be achieved by examining the evolution in shape of isocontours taken at different values of the pocket density, as shown in Figure S2.

Here the transient channel between the haem distal pocket (DP) and Xe4 is observed at higher densities. Transient channel opening is then found between sites Xe4 and Xe2, and between Xe2 and Xe3. Finally, channel opening to Xe1 is found to be the less frequent event.

It is worth noting that the pattern of migration pathways provided by MDpocket analysis agrees with the experimental findings reported for CO migration using time dependent X-ray crystallography (Schotte et al., 2003; Tomita et al., 2009). The experiments indicate that migration of CO from the distal pocket to Xe4 is an initiating event in ligand migration appearing in the nanosecond time-scale (Schotte et al., 2003). Subsequent transitions involve the migration from Xe4 to sites 2 and 3. Furthermore, those studies also indicate that CO resides in pocket Xe1 for long periods. This fact can indicate that CO is either very stable in this pocket or that migration to other pockets is unfavoured by less frequent opening of transient channels. Our results support this latter possibility.

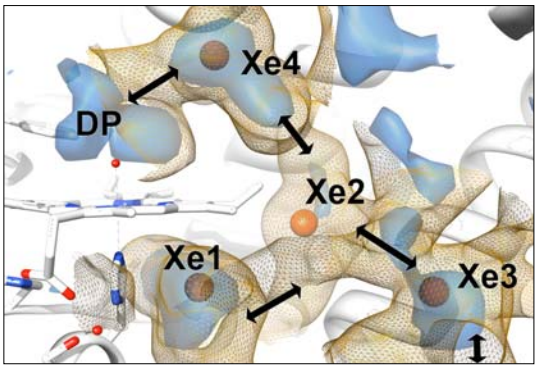


Fig. 4. Xenon atoms (orange spheres) from PDB structure 1J52 superimposed with MDpocket results on the myoglobin MD trajectory. Blue iso-surface: The pocket frequency map at 60%, which exhibits a close correspondence with all Xe binding sites. Orange iso-mesh: The pocket density map at 2 allows to discern putative migration channels from one Xe binding site to the other.

Further evidences of the transient character of channel opening are shown in Figure S3. Here pocket density maps are derived using either 10, 1000 or 10000 conformations of Mb. What can be observed is that stable pockets (red arrows) are identified even using few snapshots. However, transient channels (green arrow and blue arrow) can just be partially or even not found. Thus, usage of more conformational information allows for visualisation of these transient channels (green arrow).

The passage of diatomic ligands from the solvent into the distal haem cavity is another relevant phenomenon in globins. For Mb, it is generally accepted that HisE7 acts as gating residue for the entry pathway (Johnson, J S Olson, & G N Phillips, 1989; J S Olson et al., 1988; Perutz, 1970; E E Scott & Q H Gibson, 1997). This gating mechanism is related to a rotation around the HisE7 Cα-Cβ bond that places the residue outside the distal cavity. The opening event can be directly related to the appearance of a connecting pathway between the protein surface and the distal cavity detected with MDpocket. The frequency of appearance is thus a rough measure of the accessibility of the cavity, which in this case corresponds to around 30%. It must be noted that this value is around 10% higher than the alternate entry pathway through the Xe binding pockets. These results are in accordance with the experimental measurements that propose the HisE7 route as the main entry to the distal cavity (Emily E Scott et al., 2001).

In conclusion, MDpocket allows easy and straightforward tracking and characterization of transient internal migration channels taking advantage of the wealth of conformational space sampled in molecular dynamics.

3.5 Case 2: MDpocket as a tool for docking

Docking of small molecules into binding sites is an outstanding tool in drug discovery (Jorgensen, 2004). While the flexibility of the ligand is generally addressed in modern docking software, very often only a rigid representation of the protein target is considered. Obviously, this can have a direct impact in limiting the enrichment in virtual screening studies. Thus, efforts are being undertaken to include protein motion in molecular docking (Henzler & Rarey, 2010). Here we will show possible applications of MDpocket in ensemble docking strategies. It is known that docking results can be improved by using conformational ensembles of proteins (Barril & Morley, 2005; Novoa et al., 2010) as a strategy to provide a better resolution on the motion of the binding site. One problem in using such ensembles is the selection of conformations to be included in a systematic docking approach (Rueda, Bottegoni, & Abagyan, 2010). To test the suitability of MDpocket to guide the selection of those structures, the P38 Map kinase protein has been examined due to the known flexibility of its binding site. Thus, docking on different conformations should be considered on a target like P38 (Sherman, Beard, & Farid, 2006). In order to assess whether MDpocket can be used for efficient conformation selection for molecular docking, the ATP binding site was specifically tracked using MDpocket and characterized by means of two descriptors: the pocket volume and the mean local hydrophobic density. Whereas the volume reflects the fact that the size of the binding site is a major limitation to the binding of compounds, the second is a powerful predictor of the druggability of a binding site (Schmidtke & Barril, 2010).

During the MD trajectory run for P38 Map kinase, it can be observed that the P-loop lining the binding site is opening. This conformational rearrangement is relevant, as it allows increase of required space for docking certain P38 binders. However, a complete opening of the cavity yields a huge and very open pocket. Partially due to this opening of the loop, and furthermore the relative motion between the lobes, the pocket volume of the active site is increasing (see Figure 5). Though an open pocket can fit a larger variety of molecules without large steric clashes, a completely open binding site is not necessarily suitable to efficiently fit a given ligand.

To identify which conformations of the receptor can be theoretically considered for docking, 32 binders of P38 Map kinase with known crystallographic structure (see List S2) were superimposed to all snapshots of the MD trajectory using the beta sheet lining the binding site as reference. The ligand was then extracted from the superimposed crystal structure and inserted without altering its bioactive conformation into each snapshot of the trajectory.

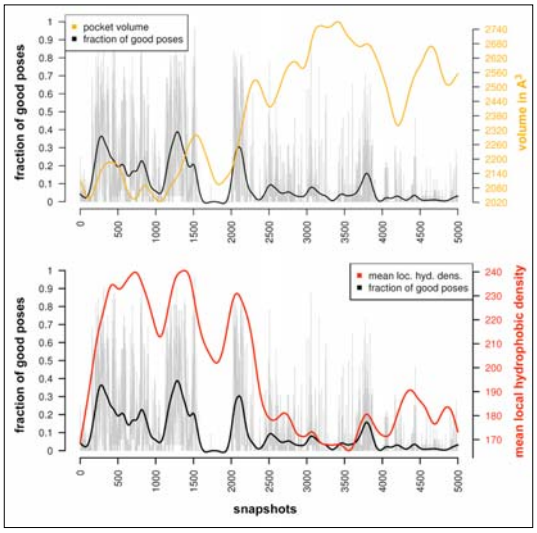


Fig. 5. Fraction of binding poses without steric clashes (grey) of 32 binders on 5000 snapshots of the trajectory run for P38 Map kinase. Smoothed values are shown in black. Pocket volume and mean local hydrophobic density of the pocket are shown in orange and red, respectively. The mean local hydrophobic density is correlated with the fraction of good binding poses.

Next, the interaction energy between the ligand and the protein was calculated. If the interaction energy is stabilizing (negative value), then the pose of the ligand (kept in the bioactive conformation) is considered to be sterically acceptable in the current MD snapshot of the protein. It should be emphasized that this computational strategy was conceived with a twofold purpose. First, it provides a simple tool to identify whether the structural features of a given snapshot are suited to accommodate the ligand in its bioactive conformation. Second, it allows us to circumvent the flaws and limitations associated to ligand sampling and scoring inherent to classical docking approaches, which might lead to prediction of poses dissimilar to the X-ray ones.

The fraction of acceptable poses out of the 32 known binders is tracked over time in Figure 5.. The upper plot shows no significant correlation between the fraction of acceptable poses and the volume of the cavity. In fact, the results indicate that although the volume is notably enlarged after the first 2000 snapshots, as reflected in the increase of the RMSD for all heavy atoms of residues lining the ATP binding site (see Figure S5), a bigger binding site is not necessarily the most suitable pocket to bind small molecules. This trend can be ascribed to the concomitant reduction of the putative interaction surface between ligand and receptor. Thus, the selection of interesting conformations suitable for molecular docking using only the volume of the binding-site as indicator is not recommended. Therefore, the mean local hydrophobic density, a descriptor that reflects local densities of hydrophobic alpha sphere clusters in a binding site and has been shown to correlate with druggability (Schmidtke & Barril, 2010) was examined. As shown in the lower plot of Figure 5, there is a striking correlation between the mean local hydrophobic density and the fraction of good poses. A tentative explanation for this good correlation is that the mean local hydrophobic density catches situations in time, where hydrophobic patches are most accessible in the binding site and the binding site is compact. Accessibility of the hydrophobic surface is likely to have a beneficial effect on binding the rather hydrophobic drug-like molecules (Vieth et al., 2004).

Overall, the preceding results suggest that tracking the mean local hydrophobic density during a MD trajectory on a pocket can give hints on the suitability of protein conformations for molecular docking. For instance, one can consider snapshots of the MD trajectory where the mean local hydrophobic density is maximised to use them as conformations for molecular docking. To the best of our knowledge, MDpocket is the first tool that might be able to identify conformations suited to bind small molecular binders, opening a variety of possible applications and rendering MD trajectories of proteins more accessible for molecular docking.

3.6 Comparison to existing methods

This section intends to propose a comparative analysis of MDpocket with other methods designed for prediction of transient pockets and channels. With regard to pockets, the comparison is limited to **EPOS^{BP}** (Eyrisch & Helms, 2007), whereas a larger variety of methods is considered for the detection of channels.

Only few tools have been reported to address the identification of transient drug binding sites. Using the PASS cavity detection algorithm (Brady & Stouten, 2000), EPOS^{BP} defines pocket volumes via active site points and neighbouring probe positions identified by PASS. Using such an active site volume, all pocket-lining atoms (PLA) are identified as being located not more than 5 Å from it. After execution of these steps on all snapshots of an MD trajectory, sets of PLAs can be compared to each other and clustered together to find conserved cavities, which are assigned a pocket Id and whose pocket properties are tracked..

A major difference with MDpocket is that pockets identified with EPOS^{BP} are mapped to a set of PLAs, while MDpocket maps pockets to a grid representation. While mapping cavities to pocket-lining atoms might provide insensitivity to rotation and translation of the protein, it hinders the usage of the method on structural ensembles other than MD trajectories (despite the attempt to use residue names and atom names). Importantly, this is especially useful when analysing homologous structures for conserved cavities. On the other hand, EPOS^{BP} allows tracking very few pocket descriptors: volume, depth and polarity., The usage of this latter descriptor might be debatable as it likely includes polar atoms in the binding pocket that do not contribute to the accessible surface area. MDpocket allows tracking a large set of pocket descriptors available via fpocket (for example pocket size, polar and apolar surface areas, hydrophobicity and polarity measures, pocket density and average radius, local hydrophobic density…). The availability of the source code of MDpocket also allows implementation of novel descriptors, but disclosure of EPOS^{BP}'s and PASS's source code hinders this.

Using our example of 88 crystal structures of HSP90, we note two major advantages of MDpocket compared to EPOS^{BP}: pocket tracking and performance. When the sub-pocket is open, the ATP binding site is identified by EPOS^{BP} as a separate pocket with a distinct pocket Id (36.4% of cases), thus making it difficult to track

large changes of cavity shape using a continuous representation. Computational performance of MDpocket is >25-fold faster making it better suited to analyze long MD trajectories; on this set, EPOS^{BP} performed pocket detection and clustering (without analysis and property tracking) in 9 minutes versus 20 seconds for MDpocket on a single core of an Intel Q9550, 2.83Ghz.

A multitude of methods exist to tentatively detect transient channels and visualize them, such as Caver/Mole, MolAxis, implicit ligand sampling (ILS) and DyME.

In **Caver** (Beneš et al., 2010) a grid is superimposed to the protein core and a starting position on this grid has to be defined. A value is associated to each grid point dependent on the radius of the biggest contact sphere that could be fit into the channel. Next a modified version of the Dijkstra shortest path detection algorithm is used to find the optimal path from the starting point inside the protein to the outside. Caver can be used on MD trajectories to identify conserved entry gorges for migration channels. However, the need to define the starting position on the grid might limit the capabilities of Caver as an exploratory tool to find transient channels.

More recently the same authors published MOLE (Petrek et al., 2007), which relies on the same algorithmic principles as Caver, but uses Voronoi tessellation of the interior of the protein to find ideal paths, has better execution speeds and allows analysis of bigger systems. However, the requirement for defining the starting point in the search of channels persists in MOLE.

MolAxis (Yaffe et al., 2008) pursues to identify migration channels for small molecules inside the protein. Typical examples are given as channels observed in transmembrane proteins and cytochrome P450. Using the alpha shape and the medial axis, MolAxis allows identifying so-called corridors or probable migration pathways of small ligands. Similarly to Caver, MolAxis requires the starting position for searching channels to be defined. In comparison to Caver, MolAxis is significantly faster and thus facilitates the analysis of larger systems and large sets of MD trajectory snapshots.

Importantly, all of the previously cited methods are able to perform channel detection on a single structure. For example MolAxis, allows tracking of channel radii during MD trajectories, but not the actual detection of transient channels as such, as the channel has to be predefined (or detected) and then tracked.

In **Implicit Ligand Sampling (ILS)** (Cohen et al., 2008), all snapshots of the MD trajectory are superimposed to a reference structure, and then a suitable probe (i.e., a gaseous ligand) is moved through a regular grid inside the protein in order to determine the interaction energy with the protein, which is finally used to detect favorable migration paths. Like MDpocket, ILS performs a complete and exploratory cavity search. Results can be visualised as PMF maps using VMD. Unlike MDpocket, ILS is conceived for (i) detection of gas migration pathways and (ii) use on MD trajectories only (i.e., need of consistent topology between different protein conformations). Last, ILS itself is a purely exploratory tool, allowing the creation of PMF maps, but not allowing their analysis or the extraction of channel properties.

DyME (Dynamic Map Ensemble) was published during the writing of this manuscript (Lin & Song, 2011). The initial steps of the DyME workflow and the need for a set of conformations are very similar to MDpocket. DyME detects all Voronoi vertices on and inside the protein and discards vertices with low clearances (radii of the alpha spheres in MDpocket). Remaining vertices are reduced to a maximum spanning tree. Next, vertices from the spanning tree are calculated for every conformation and conserved vertice clusters away from the bulk solvent (internal cavities) are identified. These clusters are then mapped back to the Voronoi vertice spanning tree of each conformation of the protein. Next putative portal regions on the surface of the protein are identified, and last a so-called super graph is computed connecting stable cavities via maximum clearance channels observed at least once in an ensemble of protein conformations between each other and portal regions.

To show putative applications of DyME the authors also used Myoglobin as example. Thus, results obtained via MDpocket can be compared with DyME. The density grid produced by MDpocket (iso-value 1.8) is shown on figure S4 in mesh representation. Results obtained from DyME are superimposed as green pathways. As MDpocket results are more exploratory and contain more information, only channels around the Xe2, S1 and S2 pockets are shown for MDpocket. Interestingly, as with DyME, MDpocket identifies previously observed sites S1 and S2 (Bossa et al., 2004). Furthermore connections of these sites to the surface gates P7, P9 and P10 can also be observed as with DyME. Although not shown explicitly, gates like P1, P2, P4 and P6 are also identified by MDpocket. The overall coverage between DyME and MDpocket results is excellent.

Both DyME and MDpocket have several advantages over previously cited channel detection methods, notably the fact that both allow identification of the complete channel network and their flexibility of application. However, DyME infers the existence of small pockets via a clustering algorithm and the existence of connections between them via at least one occurrence of usually very low radius connection channels using Voronoi edges. MDpocket uses a more generic approach based on observations where alpha spheres are found (and thus real voids) during time, how often and with what density (similar to cavity densities in DyME). Thus MDpocket results correspond to physically more meaningful connection paths, while DyME paths can be true (if the clearance is high) but might also be inferred (if the clearance is low). This fact can be clearly observed in Figure 7 of (Lin & Song, 2011), which shows the distribution of clearances for all channels for a 10 ns MD trajectory. While a significant amount of snapshots infer clearances of a radius of 1.5 Å, very few do have a clearance radius of 1.7 Å and above. Using MDpocket density and frequency maps and the corresponding detection parameters, results show actual voids of a minimum radius seen during a MD trajectory. Like DyME, MDpocket can also identify inferred channels by reducing the minimum size for alpha spheres.

An apparent advantage of DyME is the measure of the channel radii. MDpocket, however, can track various properties of the channel (or pocket) and show their time evolution along the simulation. On the other hand, the computational efficiency of MDpocket is much better. Finally, MDpocket is standalone software distributed under the GNU GPL and is not based on commercial software like MATLAB.

4 DISCUSSION

MDpocket is a new tool intended to facilitate the qualitative and quantitative analysis of transient pockets and channels from conformational ensembles that account for the intrinsic dynamics of proteins. MDpocket is based on *fpocket*, a general-purpose pocket detection program, allowing adaptable, fast, free and reliable pocket detection. The accuracy of pocket prediction is mainly limited by two factors: i) the grid-based nature of MDpocket, and ii) the choice of parameters for filtering and clustering alpha spheres.

As MDpocket uses a grid-based methodology, the results can be affected by the necessary structural superimposition of the snap-

shots. Results derived for P38 Map kinase, where the two domains exhibit large relative motion with respect to each other, show a dependence on the set of atoms used for alignment. If the alignment is done using as reference one of the two subunits, pocket detection and averaging during time will be altered for all pockets on the other subunit. Consequently, pocket frequency and density on the second subunit are underestimated. In such a case one should perform the structural alignment on the cavity of interest (the active site in our case), and bear in mind that pockets found on other places of the protein might not be representative of the whole trajectory.

The use of different parameter sets for pocket detection can have a major influence on results. Here, three parameter sets have been assessed, each addressing a given purpose. For exploratory MDpocket runs, probably water probe-sized alpha spheres (parameter set 3) are sufficient to give insights into the pocket behaviour during MD trajectories. However, if the main aim is to explore internal pathways in the protein matrix, parameter set 2 is better suited. This is illustrated by the analysis of the 50 ns MD simulation of Mb. MDpocket was capable of identifying crystallographically known Xe-binding pockets and infer putative connections between these binding sites in agreement with experimental data. Furthermore, MDpocket allows for identification of the HisE7 migration pathway, which is considered to be the main entry of diatomic ligands to the distal haem cavity.

Exploration of binding sites adapted for compounds (i.e., drugs, substrates) larger than gaseous small molecules can be simply made by changing the search parameters (default parameter set). This versatility is illustrated by "simulating" the docking of 32 ligands to P38 Map kinase. Virtual screening techniques often consider a rigid receptor conformation or at most a set of receptor conformations (ensemble docking). In this latter case, it is often very difficult to choose representative receptor conformations for molecular docking (Rueda et al., 2010). The results derived for P38 Map kinase indicate that the combined use of MD and MDpocket with a descriptor related to druggability (mean local hydrophobic density) could be valuable to identify timeframes where a pocket of interest is in a good conformation to fit a drug-like molecule. These results strongly suggest that MDpocket could be used for an easy selection of protein conformations for ensemble docking strategies.

MDpocket is a versatile tool to render mechanistic studies of proteins involving protein motion easier. The software's adaptability and scalability allows applications in various domains like conformational selection for molecular docking or the study of structural plasticity of drug binding sites. Comparing MDpocket to existing methods, it is noteworthy that the program is the first tool addressing both channel and pocket detection and also characterisation in one single framework. Contrary to most channel prediction methods, MDpocket follows an exploratory strategy that renders putative migration channels and binding sites. MDpocket is the only algorithm allowing easy selection of user defined zones and further extraction a large set of time dependent descriptors of the selected zone. Finally, MDpocket is published within the fpocket suite of pocket detection programs, under the GNU GPL License and is freely available for download on http://fpocket.sourceforge.net.

ACKNOWLEDGEMENTS

We thank Vincent Le Guilloux, Julien Maupetit and Pierre Tufféry for helpful discussions and support. We thank Allen Lin and Guang Song for providing results obtained in myoglobin with DyME. We thank the Spanish Ministerio de Ciencia e Innovación (SAF2008-0559, SAF2009-08811) and the Generalitat de Catalunya (FI fellowship to PS; grant SCG-2009-294) for financial support.

REFERENCES

Ali, M. M. U., Roe, S. M., Vaughan, C. K., Meyer, P., Panaretou, B., Piper, P. W., et al. (2006). Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature*, *440*(7087), 1013-7.

An, J., Totrov, M., & Abagyan, R. (2005). Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & cellular proteomics : MCP*, *4*(6), 752-61.

Barril, X., & Morley, S. D. (2005). Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *Journal of medicinal chemistry*, *48*(13), 4432-43. American Chemical Society.

Beneš, P., Eva, C., Barbora, K., Anton'in, P., Ond\vrej, S., Jan, B., et al. (2010). CAVER 2.1.

Bidon-Chanal, Axel, Martí, M. A., Crespo, A., Milani, M., Orozco, M., Bolognesi, M., et al. (2006). Ligand-induced dynamical regulation of NO conversion in Mycobacterium tuberculosis truncated hemoglobin-N. *Proteins*, *64*(2), 457-64.

Borrelli, K. W., Vitalis, A., Alcantara, R., & Guallar, V. (2005). PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *Journal of Chemical Theory and Computation*, *1*(6), 1304-1311. American Chemical Society.

Bossa, C., Anselmi, M., Roccatano, D., Amadei, A., Vallone, B., Brunori, M., et al. (2004). Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophysical journal*, *86*(6), 3855-62.

Brady, G. P., & Stouten, P. F. (2000). Fast prediction and visualization of protein binding pockets with PASS. *Journal of computer-aided molecular design*, *14*(4), 383-401.

Brooks, B. R., Brooks III, C. L., Mackerell Jr., A. D., Nilsson, L., Petrella, R. J., Roux, B., et al. (2009). CHARMM: The Biomolecular Simulation Program. *JOURNAL OF COMPUTATIONAL CHEMISTRY*, *30*(10, Sp. Iss. SI), 1545-1614. 111 RIVER ST, HOBOKEN, NJ 07030 USA: JOHN WILEY & SONS INC.

Carrillo, O., & Orozco, M. (2008). GRID-MD-A tool for massive simulation of protein channels. *Proteins*, *70*(3), 892-9.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., et al. (2005). The Amber biomolecular simulation programs. *Journal of computational chemistry*, *26*(16), 1668-88.

Chemical Computing Group. (2009). MOE (The Molecular Operating Environment) Version 2009.10.

Cohen, J., Arkhipov, A., Braun, R., & Schulten, Klaus. (2006). Imaging the migration pathways for O2, CO, NO, and Xe inside myoglobin. *Biophysical journal*, *91*(5), 1844-57.

Cohen, J., Olsen, K. W., & Schulten, Klaus. (2008). Finding gas migration pathways in proteins using implicit ligand sampling. *Methods in enzymology*, *437*, 439-57.

Eyrisch, S., & Helms, V. (2007). Transient pockets on protein surfaces involved in protein-protein interaction. *Journal of medicinal chemistry*, *50*(15), 3457-64.

Glazer, D. S., Radmer, R. J., & Altman, R. B. (2009). Improving structure-based function prediction using molecular dynamics. *Structure (London, England : 1993)*, *17*(7), 919-29.

Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. *Chemical biology & drug design*, *69*(2), 146-8.

Hendlich, M., Rippmann, F., & Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of molecular graphics & modelling*, *15*(6), 359-63, 389.

Henrich, S., Salo-Ahen, O. M. H., Huang, B., Rippmann, F. F., Cruciani, G., & Wade, R. C. (2010). Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of molecular recognition : JMR*, *23*(2), 209-19.

Henzler, A. M., & Rarey, M. (2010). In Pursuit of Fully Flexible Protein-Ligand Docking: Modeling the Bilateral Mechanism of Binding. *Molecular Informatics*, *29*(3), 164-173.

Hess, B., Kutzner, C., Spoel, D. van der, & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular

Simulation. *Journal of Chemical Theory and Computation*, *4*(3), 435-447. American Chemical Society.
Huang, B. (2009). MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omics : a journal of integrative biology*, *13*(4), 325-30.
Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33-8, 27-8.
Johnson, K. A., Olson, J S, & Phillips, G N. (1989). Structure of myoglobin-ethyl isocyanide. Histidine as a swinging door for ligand entry. *Journal of molecular biology*, *207*(2), 459-63.
Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science (New York, N.Y.)*, *303*(5665), 1813-8.
Kim, D., Cho, C.-H., Cho, Y., Ryu, J., Bhak, J., & Kim, D.-S. (2008). Pocket extraction on proteins via the Voronoi diagram of spheres. *Journal of molecular graphics & modelling*, *26*(7), 1104-12.
Kleywegt, G. J., & Jones, T. A. (1994). Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta crystallographica. Section D, Biological crystallography*, *50*(Pt 2), 178-85.
Kua, J., Zhang, Y., & McCammon, J. A. (2002). Studying enzyme binding specificity in acetylcholinesterase using a combined molecular dynamics and multiple docking approach. *Journal of the American Chemical Society*, *124*(28), 8260-7.
Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, *13*(5), 323-30, 307-8.
Laurie, A. T. R., & Jackson, R. M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)*, *21*(9), 1908-16.
Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, *10*, 168.
Leis, S., Schneider, S., & Zacharias, M. (2010). In silico prediction of binding sites on proteins. *Current medicinal chemistry*, *17*(15), 1550-62.
Liang, J., Edelsbrunner, H., & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein science : a publication of the Protein Society*, *7*(9), 1884-97.
Lin, T.-L., & Song, G. (2011). Efficient mapping of ligand migration channel networks in dynamic proteins. *Proteins*, *79*(8), 2475-90.
MacKerel Jr., A. D., Brooks III, C. L., Nilsson, L., Roux, B., Won, Y., & Karplus, M. (1998). CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In P. v. R. Schleyer et al. (Ed.), (Vol. 1, pp. 271-277). John Wiley & Sons: Chichester.
Marti, M. A., Crespo, A., Capece, L., Boechi, L., Bikiel, D. E., Scherlis, D. A., et al. (2006). Dioxygen affinity in heme proteins investigated by computer simulation. *Journal of inorganic biochemistry*, *100*(4), 761-70.
Novoa, E. M., Pouplana, L. R. de, Barril, X., & Orozco, M. (2010). Ensemble Docking from Homology Models. *Journal of Chemical Theory and Computation*, *6*(8), 2547-2557. American Chemical Society.
Olson, J S, Mathews, A. J., Rohlfs, R. J., Springer, B. A., Egeberg, K. D., Sligar, S. G., et al. (1988). The role of the distal histidine in myoglobin and haemoglobin. *Nature*, *336*(6196), 265-6.
Ostermann, A., Waschipky, R., Parak, F. G., & Nienhaus, G. U. (2000). Ligand binding and conformational motions in myoglobin. *Nature*, *404*(6774), 205-8. Macmillian Magazines Ltd.
Pérot, S., Sperandio O., Miteva, M. A., Camproux, A.C. & Villoutreix, B.O. (2010). Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, 15(15-16), 656-667
Perutz, M. F. (1970). Stereochemistry of cooperative effects in haemoglobin. *Nature*, *228*(5273), 726-39.
Peters, K. P., Fauck, J., & Frömmel, C. (1996). The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *Journal of molecular biology*, *256*(1), 201-13.
Petrek, M., Kosinová, P., Koca, J., & Otyepka, M. (2007). MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure (London, England : 1993)*, *15*(11), 1357-63.
Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605-12.
Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, *26*(16), 1781-802.
Rueda, M., Bottegoni, G., & Abagyan, R. (2010). Recipes for the selection of experimental protein conformations for virtual screening. *Journal of Chemical Information and Modeling*, *50*(1), 186-193. ACS Publications.
Schmidtke, P., & Barril, X. (2010). Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of medicinal chemistry*, *53*(15), 5858-67.
Schotte, F., Lim, M., Jackson, T. A., Smirnov, A. V., Soman, J., Olson, John S, et al. (2003). Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science (New York, N.Y.)*, *300*(5627), 1944-7.
Scott, E E, & Gibson, Q H. (1997). Ligand migration in sperm whale myoglobin. *Biochemistry*, *36*(39), 11909-17.
Scott, Emily E, Gibson, Quentin H, & Olson, John S. (2001). Mapping the Pathways for O2 Entry Into and Exit from Myoglobin. *Journal of Biological Chemistry*, *276*(7), 5177-5188.
Seco, J., Luque, F. J., & Barril, X. (2009). Binding site detection and druggability index from first principles. *Journal of medicinal chemistry*, *52*(8), 2363-71. American Chemical Society.
Sherman, W., Beard, H. S., & Farid, R. (2006). Use of an induced fit receptor structure in virtual screening. *Chemical biology & drug design*, *67*(1), 83-4.
Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., & Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. *Journal of medicinal chemistry*, *49*(2), 534-53. American Chemical Society.
Spyrakis, F., Faggiano, S., Abbruzzetti, S., Dominici, P., Cacciatori, E., Astegno, A., et al. (2011). Histidine E7 dynamics modulates ligand exchange between distal pocket and solvent in AHb1 from Arabidopsis thaliana. *The journal of physical chemistry. B*, *115*(14), 4138-46.
Tilton, R. F., Kuntz, I. D., & Petsko, G. A. (1984). Cavities in proteins: structure of a metmyoglobin xenon complex solved to 1.9 .ANG. *Biochemistry*, *23*(13), 2849-2857. American Chemical Society.
Tomita, A., Sato, T., Ichiyanagi, K., Nozawa, S., Ichikawa, H., Chollet, M., et al. (2009). Visualizing breathing motion of internal cavities in concert with ligand migration in myoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(8), 2612-6.
Vieth, M., Siegel, M. G., Higgs, R. E., Watson, I. A., Robertson, D. H., Savin, K. A., et al. (2004). Characteristic physical properties and structural fragments of marketed oral drugs. *Journal of medicinal chemistry*, *47*(1), 224-32. American Chemical Society.
Weisel, M., Proschak, E., & Schneider, G. (2007). PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central journal*, *1*(1), 7.
Wright, L., Barril, X., Dymock, B., Sheridan, L., Surgenor, A., Beswick, M., et al. (2004). Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chemistry & biology*, *11*(6), 775-85.
Yaffe, E., Fishelovitch, D., Wolfson, H. J., Halperin, D., & Nussinov, R. (2008). MolAxis: efficient and accurate identification of channels in macromolecules. *Proteins*, *73*(1), 72-86.
Yang, F., & Phillips, G N. (1996). Crystal structures of CO-, deoxy- and metmyoglobins at various pH values. *Journal of molecular biology*, *256*(4), 762-74.

Supplementary Information

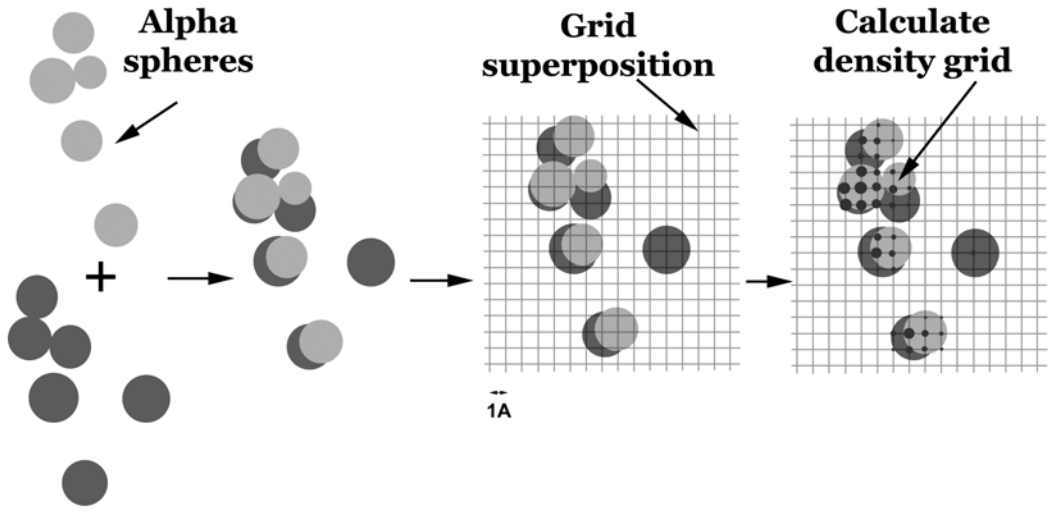
List S1:

1BYQ, 1OSF, 1UY6, 1UY7, 1UY8, 1UY9, 1UYC, 1UYD, 1UYE, 1UYF, 1UYG, 1UYH, 1UYI, 1UYK, 1UYL, 1UYM, 1YC1, 1YC3, 1YC4, 1YER, 1YES, 1YET, 2BSM, 2BT0, 2BYH, 2BYI, 2BZ5, 2CCS, 2CCT, 2CCU, 2CDD, 2FWY, 2FWZ, 2H55, 2JJC, 2K5B, 2QF6, 2QFO, 2QG0, 2QG2, 2UWD, 2VCI, 2VCJ, 2WI1, 2WI2, 2WI3, 2WI4, 2WI5, 2WI6, 2WI7, 2XAB, 2XDK, 2XDL, 2XDS, 2XDU, 2XDX, 2XHR, 2XHT, 2XHX, 2XJG, 2XJJ, 2XJX, 2XK2, 3BM9, 3BMY, 3D0B, 3EKO, 3EKR, 3FT5, 3FT8, 3HEK, 3HHU, 3HYZ, 3HZ1, 3HZ5, 3INW, 3INX, 3K97, 3K98, 3K99, 3MNR, 3NMQ, 3QTF, 3R4M, 3R4N, 3R4O, 3R4P

List S2 :

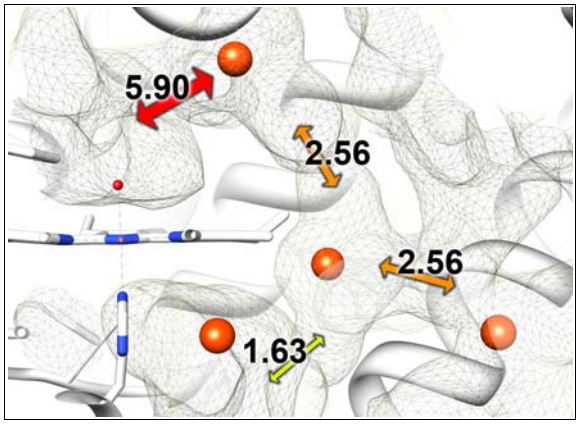
1M7Q, 1OUK, 1OUY, 1ZYJ, 1ZZ2, 1ZZL, 2BAL, 2GTM, 2GTN, 2QD9, 2RG6, 2ZAZ, 3FC1, 3FKO, 3FL4, 3FLQ, 3FLZ, 3FI4, 3FL4, 3FMH, 3FMJ, 3FMK, 3FML, 3FMM, 3FSF, 3GC7, 3GCP, 3HP2, 3HRB, 3IW5, 3IW6, 3KF7

Figure S1 :



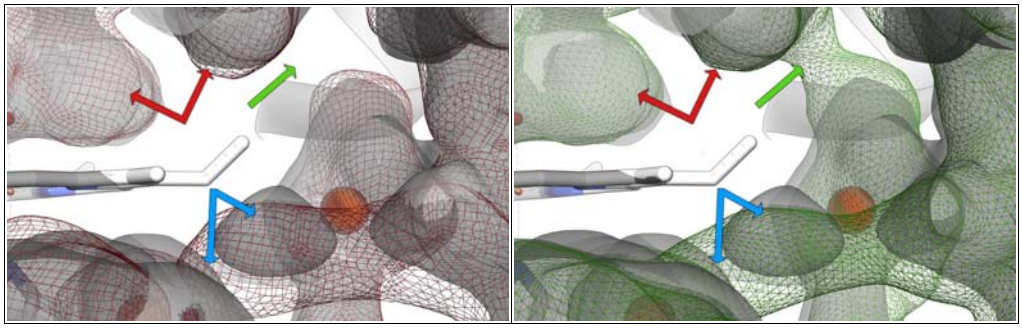
Schematic representation of the MDpocket workflow. Alpha spheres are detected on different pre-aligned conformations of the protein (dark grey 1 conformation, light grey another conformation). A 1Å spaced grid is superimposed to the alpha spheres and on each grid point the density of sourounding alpha spheres and frequency are tracked.

Figure S2 :



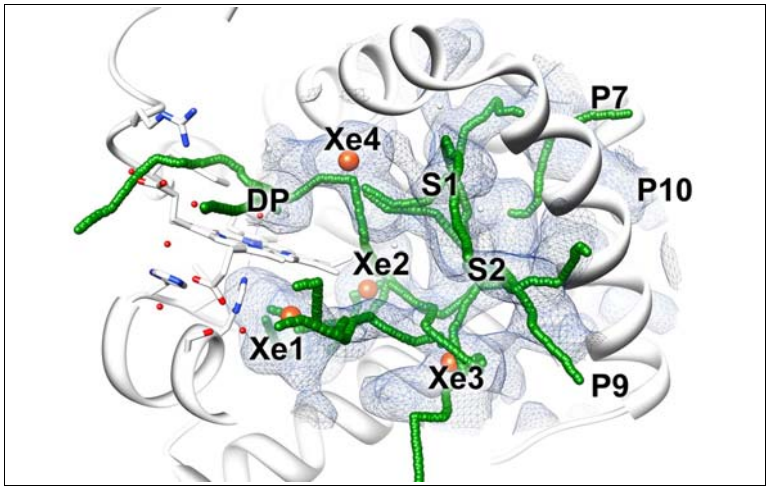
Pocket density map of myoglobin with iso-value thresholds of channel opening. Higher values infer more stable opening.

Figure S3 :



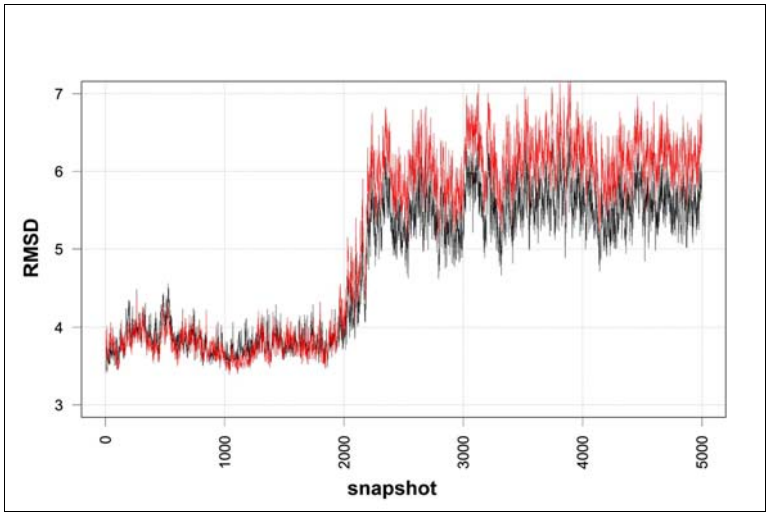
Pocket density maps of the myoglobin trajectory with different amount of snapshots used as input for MDpocket at an iso-value of 2.3. Grey iso-surface : 10 snapshots, Red iso-mesh : 1000 snapshots, Green iso-mesh : 10000 snapshots. Stable cavities (red arrows) are detected with few snapshots already. Transient channels (blue arrows and green arrow) can only be spotted by using a larger amount of snapshots.

Figure S4 :



Myoglobin transient channel networks. (Green pathways) DyME results obtained from from (Lin and Song, 2011) superimposed with results produced by MDpocket. Both procedures identify the main xenon binding sites. MDpocket also spots uncharacterized pockets S1 and S2 and entry channels P7, 10 and 9. Only partial results are shown for MDpocket for clarity.

Figure S5:



RMSD of heavy atoms of all residues lining the ATP binding site of P38 Map kinase. (black): RMSD for crystal structure 1m7q superimposed to every snapshot of a MD trajectory, (red): RMSD for crystal structure 1ouk superimposed to every snapshot of a MD trajectory. Superimposition was done using the stable beta-sheet lining the binding site as reference, but the RMSD is calculated on the whole binding site, to reproduce as accurately data used for MDpocket input.

fpocket: herramientas en línea para la detección de sitios de uniones en proteínas y el seguimiento conjunto

Peter Schmidtke, Vincent Le Guilloux, Julien Maupetit y Pierre Tufféry
Nucleic Acids Research, 2010, 1(38) - Web Server issue

La detección computación de los sitios de union de moleculas pequeñas tiene diferentes aplicaciones en el campo de la biomedicina. De notable interés son la identificación de cavidades basadas en la estructura, o apuntes funcionales en estructuras. fpocket es un programa de detección de moleculas pequeñas, cuya teoría recae en la geometría de una esfera. El servidor web de fpocket permite: (i) la detección de huecos candidatos- fpocket, (ii) seguimiento del hueco molecular durante la dinámica molecular, a fin de proporcionar información sobre la dinámica del hueco molecular- mdpocket, y (iii) una transposición para el análisis combinado de homólogos estructurales- hpocket. Esas herramientas complementarias en línea, permite abordar varias cuestiones relacionadas con la identificación y anotaciones de sitios funcionales y alostéricos, huecos transitorios y preservación del bolsillo en la evolución structural de la familia. El servidor y la documentación están disponibles de manera gratuita en: http://bioserv.rpbs.univ-paris-diderot.fr/fpocket.

Nucleic Acids Research Advance Access published May 16, 2010

Nucleic Acids Research, 2010, 1–8
doi:10.1093/nar/gkq383

fpocket: online tools for protein ensemble pocket detection and tracking

Peter Schmidtke[1], Vincent Le Guilloux[2], Julien Maupetit[3] and Pierre Tufféry[3,*]

[1]Departament de Fisicoquimica and Institut de Biomedicina (IBUB), Facultat de Farmacia, Universitat de Barcelona, 08028, Barcelona, Spain, [2]ICOA, Institut de chimie organique et analytique, University of Orleans, Orleans and [3]MTi, INSERM UMR-S973 and RPBS, Université Paris Diderot – Paris 7, F75205 Paris, France

Received February 13, 2010; Revised April 17, 2010; Accepted April 27, 2010

ABSTRACT

Computational small-molecule binding site detection has several important applications in the biomedical field. Notable interests are the identification of cavities for structure-based drug discovery or functional annotation of structures. fpocket is a small-molecule pocket detection program, relying on the geometric α-sphere theory. The fpocket web server allows: (i) candidate pocket detection—fpocket; (ii) pocket tracking during molecular dynamics, in order to provide insights into pocket dynamics—mdpocket; and (iii) a transposition of mdpocket to the combined analysis of homologous structures—hpocket. These complementary online tools allow to tackle various questions related to the identification and annotation of functional and allosteric sites, transient pockets and pocket preservation within evolution of structural families. The server and documentation are freely available at http://bioserv.rpbs.univ-paris-diderot.fr/fpocket.

INTRODUCTION

The prediction of functional sites including ligand binding sites or catalytic sites can guide the design of small molecules that could interact with a protein and modulate its function or drive the selection of targeted mutations for protein engineering. It largely relies on the identification and characterization of clefts and cavities in protein structures.

In the past two decades, various approaches have been proposed to the identification of small-molecule binding sites. These encompass geometric analysis of protein surface such as (1–9) see (10) for more references, energy calculations (11,12), the combination of these with information derived from sequences such as residue conservation (13–15), or even meta-methods combining several such approaches to improve binding site prediction (16). Over the last years, however, several new considerations have become of interest. First, the static view of protein pockets is approximative, as is the identification of these in a static image. Differences in the pocket shape and conformation between the apo and holo proteins are known for several proteins, see for example HSP90 (17) or P38 MAP kinase (18). Whether these changes are induced by the ligand or through self-conformational changes is still controversial [e.g. (19,20)]. Last, transient pockets are known to occur on protein surfaces involved in protein–protein interactions (21,22). Current pocket detection approaches provide useful tools for identifying static pockets on static snapshots provided by the Protein Data Bank (PDB). However, very few attempts (21,22), have been made to treat information available in structural families and/or derived from molecular dynamics. Finally, escaping pocket identification from experimental structures alone is also a concern in a context of intensive genome sequencing (23).

Several online services have been proposed for pocket detection such as Q-SiteFinder (11), LIGSITE^{csc}(13), CASTp (24), SCREEN (6), PocketDepth (25) or Metapocket (16). These will usually take as input a protein structure and return one or several candidate pockets. In addition to pocket detection, SplitPocket (26) and fPOP (27) provide means of functional inference by comparing the identified patches with those identified over the complete set of PDB structures.

Recently, fpocket, a new program suite allowing pocket detection was introduced (10). The method makes use of Voronoi tesselation and α-spheres to analyse the protein surface. In a reference test set, 94 and 92% of known binding pockets were correctly identified within the best three ranked pockets from the holo and apo proteins. Here, in addition to making fpocket available online, new directions for pocket detection and analysis are proposed.

The first ambition is to analyse pocket dynamics through iterative pocket tracking on a set of PDB snapshots representing various conformational states of the protein of interest. Such an approach allows one to tackle aspects such as pocket flexibility and transientness,

*To whom correspondence should be addressed. Tel: +33 1 57278374; Fax: +33 1 57278372; Email: pierre.tuffery@univ-paris-diderot.fr

channel opening or ligand-induced conformational changes. The second purpose is to explore cavity conservation among structural families, to identify potential common structural regions of interest. Both aspects are solved using a common grid-based pocket tracking approach over collection of structures.

CONCEPTS AND METHODS

fpocket

fpocket relies on the concept of α-spheres, a concept initiated by Liang and Edelsbrunner (3) and is also used by Chemical Computing Group in the SiteFinder software (http://www.chemcomp.com/). An α-sphere is a sphere in contact with four atoms on its boundary, not containing any internal atom inside. For a protein, very small spheres are located within the protein, whereas large spheres are located at the exterior. Clefts and cavities correspond to spheres of intermediate radii. Thus, it is possible to filter the ensemble of α-spheres defined from the atoms of a protein according to some minimal and maximal radii values in order to address pocket detection. Based on this, we have recently introduced the fpocket package for pocket detection. For more information refer to (10).

Pocket tracking over collection of structural frames

Given a collection of comparable protein structures, such as provided by molecular dynamics or by homology search, one challenge is to track the persistence of pockets within this set of conformations or frames. The approach used can be summarized as an iterative run of fpocket on each frame, followed by a post-analysis using a grid-based approach, as illustrated Figure 1.

In more detail, a 1 Å spaced grid is generated to encompass previously aligned conformers. The grid allows tracking of pockets (α-spheres) in very precise zones over time. On each grid point the α-sphere density of 8Å^3 volume around it is calculated, corresponding to a small box of a 2 Å sized edge. Furthermore, the associated pocket score for each α-sphere near a grid point is tracked following formula (3).

Formulas (1) and (3) describe how the densities and scores are calculated.

$$\text{density}(x,y,z) = \frac{1}{n}\sum_{1}^{n} f_{\alpha\text{-spheres}}(x,y,z), \text{ with} \quad (1)$$

$$f_{\alpha\text{-spheres}}(x,y,z) = \text{card}(\alpha\text{-spheres}_{xyz} \cap \{(x,y,z) \pm 1.0\}) \quad (2)$$

$$\text{score}(x,y,z) = \frac{1}{n}\sum_{1}^{n} g_{\alpha\text{-spheres}}(x,y,z), \text{ with} \quad (3)$$

$$g_{\alpha\text{-spheres}}(x,y,z) = \sum \text{pocket}_{\text{score}}(\{\alpha\text{-spheres}_{xyz} \cap \{(x,y,z) \pm 1.0\}\}) \quad (4)$$

(x, y, z) is a given position in the grid space and n is the number of conformations analysed.

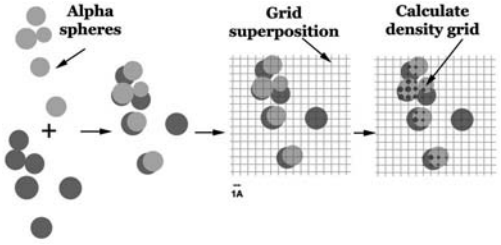


Figure 1. Workflow of the pocket tracking methodology. α-spheres from different snapshots are represented by different colors (dark and light).

The resulting grid densities can be used to analyse void space and putative migration pathways of small molecules, whereas scores can rather be used to identify conserved cavities that may bind small molecules. Similar to electron density maps, these pocket density grids can be visualized as iso-volumes, where a given isovalue v allows depiction of all grid points having a density equal to or higher than v.

The interpretation of grid density is often a complex task. Thus, the pocket grid density is mapped to a given reference protein structure using:

$$\text{density}_{\text{score}}(\text{atom}) = \frac{1}{p}\sum_{g_a t_2} \text{score}(\text{grid}_i) \quad (5)$$

where $g_a t_2$ corresponds to the grid points at a distance <2 Å of the atom and p is the number of grid points verifying this condition.

Finally, for visualization purposes using common molecular visualization tools, like PyMol (http://www.pymol.org) or VMD (28), the previously calculated density score for each atom [formula 5] is treated to match a b-factor-like float scale, allowing easy colouring. The expression used to calculate this range of values is:

$$\text{color}(\text{atom}) = \log(1+\frac{1}{2}\text{density}_{\text{score}}(\text{atom})) \quad (6)$$

mdpocket

mdpocket is the application of the pocket tracking approach to molecular dynamics trajectories. From our experience, we recommend to consider at least 200 snapshots. It can be run in two modes. The first identifies conserved as well as transient pockets and maps them to a pocket density grid. Pocket transientness can be observed by the presence of pockets of lower density in the pocket density grid, whereas high density regions correspond to stable cavities. The second mode requires a user-defined selection of grid points of interest extracted from a previous mdpocket run. It allows to focus on some specific regions of a structure. Tracking of pocket properties on grid selection is performed by considering all α-spheres within the neighbourhood of a selected grid point and merging those into one single pocket. Then all

fpocket descriptors are calculated for this pocket for each frame.

hpocket

hpocket is the application of the pocket tracking approach to collections of homologous structures. Homologous structures are identified either from sequence or from structures. Sequence-based identification is performed using CS-Blast (29) on the PDB (30), filtering the hits in terms of *e*-value, coverage, identity and maximal number. Structure-based identification makes use of the Astral/SCOP (31) classification, using the family level. The hits are then superimposed using an ancillary facility based on TM-align (32).

INPUT/OUTPUT

For all programs of the fpocket suite, two different user interfaces are provided: a classical Common Gateway Interface (CGI) (called *default* interface on the server) and a Mobyle portal (33) interface (called *advanced* interface on the server). In all cases, the same command line is called and generated results are strictly identical. Some advantages of using the Mobyle interface are: (i) the possibility (not mandatory) to open a user session by registering, which allows data persistence on the server; (ii) the possibility to bookmark data for further use, which for mdpocket can avoid the re-upload of possiblity large files; and (iii) results could be directly piped as input to other analysis programs (mdpocket Mode 1 to Mode 2 for example).

fpocket

As input, the fpocket server accepts, a simple standard PDB file or concatenated PDB files to iterate on (each file must start with the `HEADER` PDB field and ends with the `END` PDB field). On program termination, for each target, the server returns the results of the stand-alone fpocket program (10), e.g. (i) PyMol and VMD pocket visualization scripts; (ii) the query structure with embedded centers of pocket α-spheres; and (iii) each pocket (set of α-sphere centers) in a PQR file (this modified PDB format allows to set atom van der Waals radius explicitely to determine more precisely the volume detected by fpocket). Additionaly, the server provides a set of six snapshots (six sides of a cubic box) showing localization of detected pockets (Figure 2A). Moreover, predicted pockets can be downloaded independently and/or visualized through the embedded Jmol applet (http://jmol.sourceforge.net/) and the OpenAstexViewer (http://www.openastexviewer.net/) for quick analysis of the results (Figure 2B).

mdpocket

The mdpocket server requires a set of PDB snapshots to run the first step of the analysis (Mode 1). At the end of the job, the server proposes three output files: (i) the mdpocket grid that stores density information for each grid point; (ii) the pocket grid points at a particular isovalue (default is 3, i.e. grid points having 3 or more Voronoi vertices in the 8 $Å^3$ volume around the grid point for each snapshot); and (iii) the pocket α-sphere density stored in the b-factor PDB field of the first snapshot. This last file allows to quickly detect regions of interest. mdpocket allows to run a second step (Mode 2) in order to track descriptors evolution of a user-selected pocket region. To do so, the user should provide selected grid points and the previous set of PDB snapshots. As result of running mdpocket in Mode 2, frame-dependent descriptors are provided as downloadable text file, allowing further treatment using spreadsheet or statistical software like R (34). Finally, mdpocket also provides series of pictures giving an overview of the superposed structure conformational space (see legend of Figure 3A) and pocket α-sphere density (see legend of Figure 3B). Similarily to fpocket, the Jmol applet is embedded in the mdpocket results page for a quick analysis of the pockets dynamics (Figure 3C).

hpocket

hpocket provides results similarly to mdpocket with an additional homology search report, data containing the blast report, PDB hits sequences alignment and superimposed structures.

ONLINE MDPOCKET USE CASE

In the following, we detail a use case utilizing mdpocket and more briefly another more classical use case. Similar analyses could be performed in a homology context using hpocket.

A prototype of mdpocket was used to produce results published in (35). Using the same molecular dynamics for the penta-deoxy FB10L mutant of Type 1 non-symbiotic hemoglobin, Ahb1, from *Arabidopsis thaliana*, 17 snapshots equally spaced in time were submitted to the mdpocket server. Within seconds, the result depicted on the Figure 4A is produced (Mode 1). Comparing results obtained by the server with those published in (35) (Figure 4B), one can notice that the main results can be reproduced even using a very low number of snapshots compared to the number used in (35), 17 versus 800.

For this particular mutant it was shown in (35) that the geminate rebinding of carbon monoxide (CO) is not altered compared to the wild-type. Interestingly, the exit path of CO is closed (Figure 5), which alters geminate rebinding rate. However, the presence of the cavity seen beneath the heme, which is not observed in the wild-type, can explain retention of CO within the structure, without altering geminate rebinding rates as it can be seen for the HE7L mutant. This cavity is regularly disconnected from the *trans*-location pathway of CO, which is furthermore in concordance with a slower geminate rebinding rate compared to the HE7L mutant. Thus, mdpocket allows rapid evaluation of a bunch of snapshots from a Molecular Dynamics (MD) trajectory for conserved pockets or pockets that appear upon time.

As example for the mdpocket Mode 2, Ahb1 is taken again, this time using 267 snapshots picked up during the whole trajectory. Here mdpocket is used to track the volume of the connection between the pocket above the

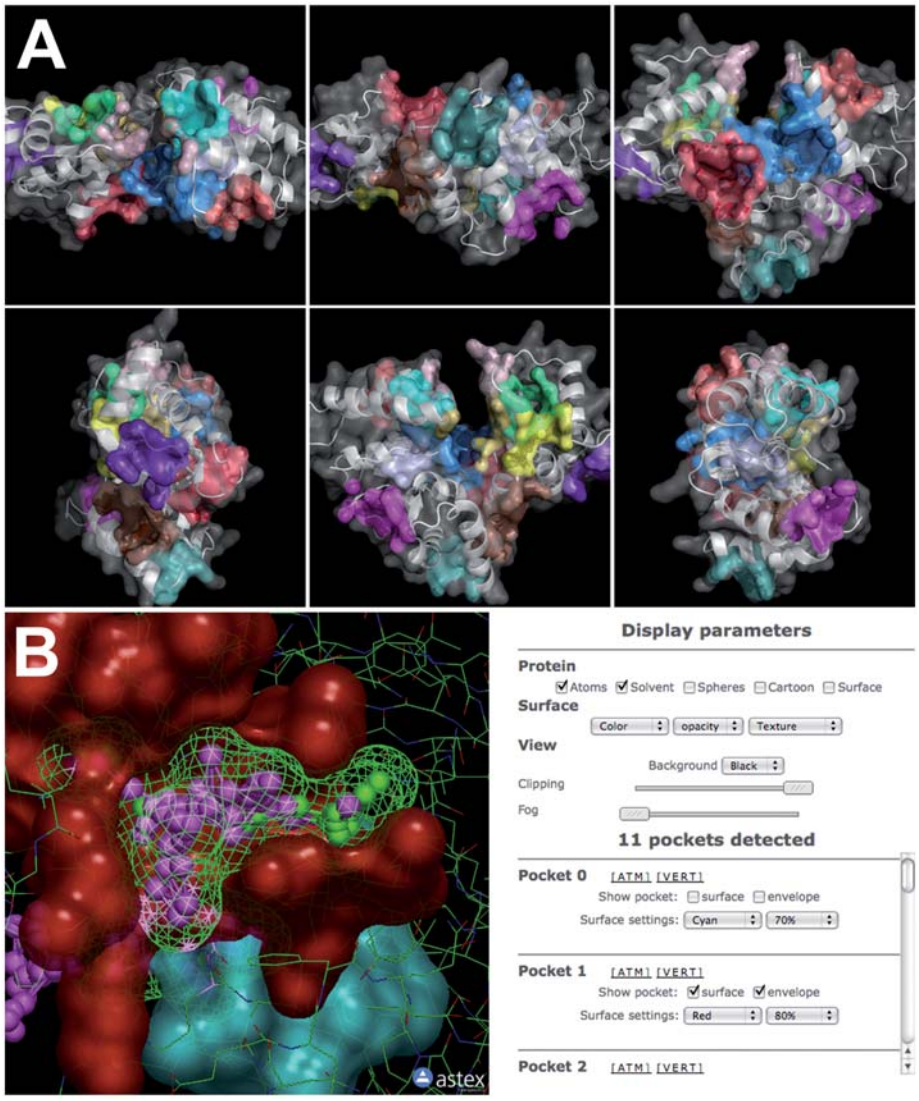


Figure 2. (**A**) fpocket server provides a set of target structure pictures, showing predicted pockets as surfaces of each pockets atoms (one color per pocket). (**B**) The fpocket results page embeds both Jmol and OpenAstexViewer (used here) applets for a quick analysis of the predicted pockets. A control panel on the right part allows the selection of the pockets to visualize and switch between various molecular representations. Here, the surface of the residues of the pocket is in red, the α-spheres in magenta and green and the envelope of the pocket is represented using a green mesh.

heme group and the cavity below the heme group. In order to do so, the selected grid points defining the pocket and the 267 snapshots were uploaded to the mdpocket server. Providing the user-defined pocket definition, mdpocket automatically tracks pocket descriptors over time on this precise zone. The resulting descriptor text file provided by the server was then used to produce the pocket volume curve in Figure 5C. Using the user-defined pocket volume, tracking of pocket descriptors intends to show transient opening/closing of this channel. A depiction of this zone is available on Figure 5A and B. The heme is situated on the left and the selected pocket grid is represented as red spheres. Furthermore, Y145 is shown in sticks as it plays an interesting role here. This run of mdpocket uses the same input data as the previous, more exploratory phase, just with the additional pocket definition. On Figure 5C, the smoothed volume (black) is plotted over time and despite fluctuations (gray), a mean volume

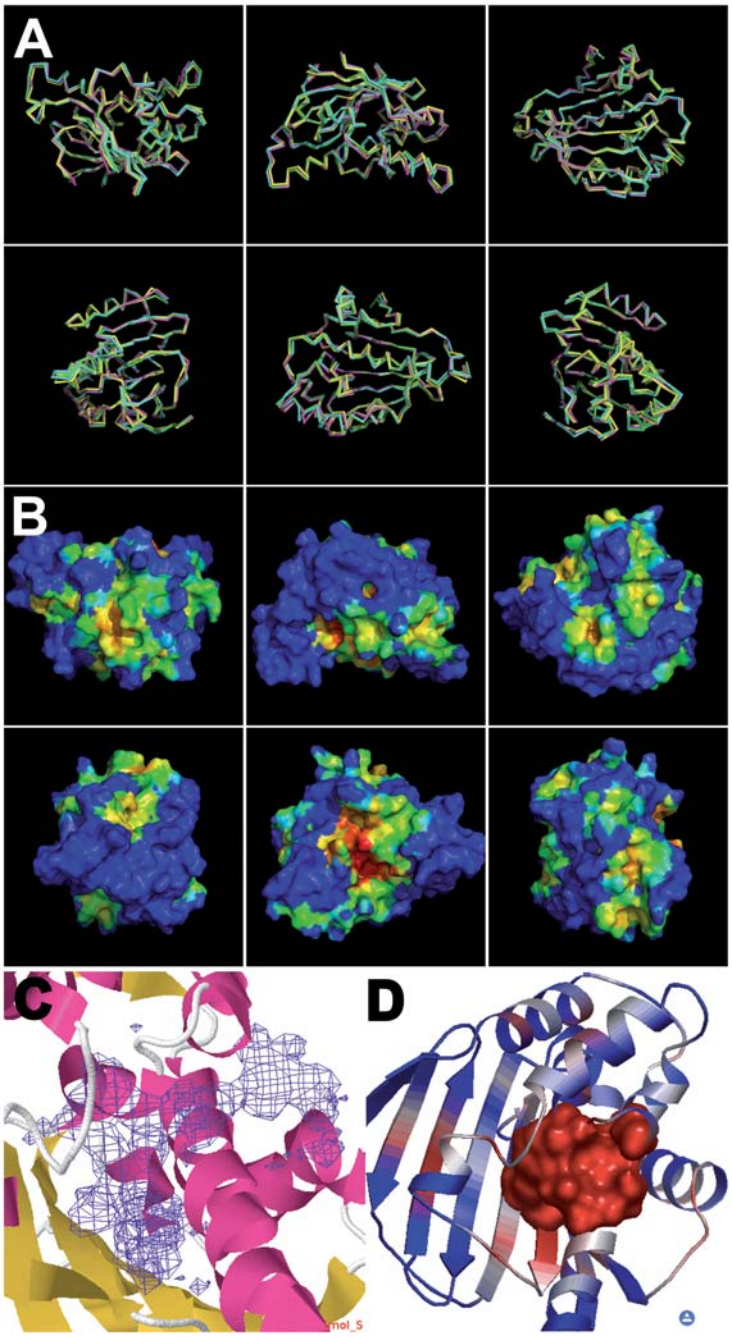


Figure 3. The mdpocket server provides a set of target structure pictures, showing superimposed PDB snapshots as ribbon (**A**) and the first snapshot structure molecular surface colored by α-spheres density ranging from blue (low density) to red (high density) (**B**). The mdpocket results page embeds the Jmol applet (**C**) to give an overview of the conserved cavities (the density grid is represented as an isosurface), and the right part provides viewing components and to extract cavity conservation at a user-selected isovalue—for more details see the main text. It is possible (**D**) to map the density information onto the residues to explore pocket stability in the structure. Here, residues corresponding to high-density regions are displayed as molecular surfaces in the OpenAstexViewer.

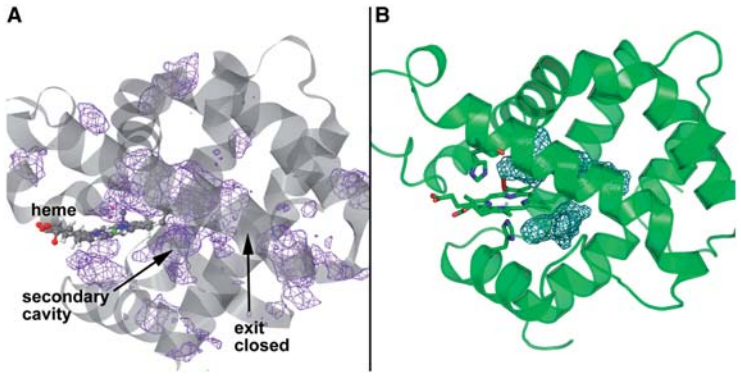


Figure 4. (**A**) Result of a run in Mode 1 of mdpocket on 17 snapshots of a MD trajectory of Ahb1 FB10L. A pocket density grid is provided, allowing visualization of conserved pockets and open channels during the MD trajectory. Analysis shows that the exit pathway of CO is closed in this mutant and that a secondary cavity beneath the heme group is existing. (**B**) Illustration of the results obtained in (35).

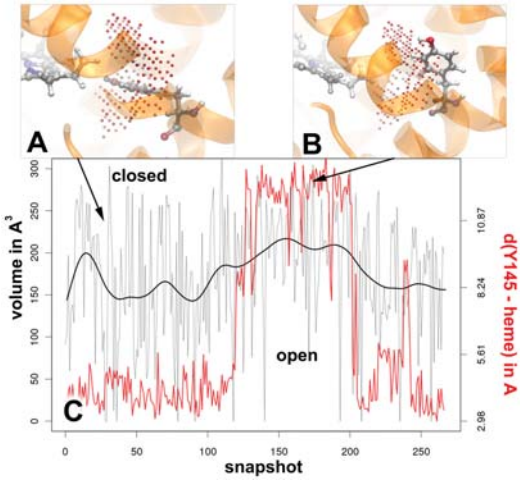


Figure 5. Tracking the volume of the channel between the upper heme pocket and lower heme pocket (mdpocket Mode 2). The pocket grid is shown as red spheres. (**A**) Y145 in the closed state is situated in the pocket, (**B**) Y145 in the open state is situated on the edge pocket. (**C**) smoothed volume of the pocket (black curve) versus time, the distance between the hydroxyl group of Y145 and the heme versus time is represented in red.

increase after 100 snapshots can be noticed. The residue Y145 is situated directly in the selected pocket (Figure 5A) and thus its position towards the pocket is measured using the distance between its hydroxyl group and the proximate heme. One can notice on Figure 5C that the volume increase corresponds to a flipping of Y145 on the side of the cavity (like that seen on figure 5B).

The small volume variation observed despite such important change in torsion angles of Y145 could be explained by the fact that Y145 is still bordering the pocket, but not obstructing it anymore. These results confirm the hypothesis that the connection between the two parts of the pocket bordering the heme is sometimes closed. Here, such a closed state can be seen during a long part of the trajectory, preceding a transient opening, which leads again to the closed state.

In the Ahb1 use case, we have illustrated how to investigate very precisely one tiny cavity. Figure 6 on P38 Map kinase shows another use case focusing on the volume of the binding site. Out of a 50 ns trajectory with explicit solvent, 1000 equally spaced snapshots were uploaded to the mdpocket server as well as the previously defined binding site. Several descriptors were tracked for the binding site of P38. Looking at the variation of the binding site volume during the trajectory, an increase of the pocket volume to 1000 $Å^3$ at around snapshot 200 can be seen. Afterwards, the mean volume of the pocket decreases to 600 $Å^3$. This very simple example shows how open/closed conformations of cavities can be isolated during MD trajectories, which can have important implications on choosing representative MD snapshots for ensemble docking, for example.

DISCUSSION AND FUTURE WORK

The fpocket web server provides a valuable, fast, free and easy to use online service allowing to tackle various aspects of protein pocket detection. It relies on a fast and efficient approach for pocket detection from a single protein structure. Furthermore, it allows to investigate new directions to explore and analyse structure ensembles.

Pocket tracking capabilities of mdpocket were shown to coincide with experimental results obtained in (35), thus the mdpocket server provides an easy interface to a new methodology for studying pocket conservation and transientness during molecular dynamics trajectories.

Due to the high scalability of the methodology behind mdpocket, it can also be used to assess pocket conservation among structural families. This functionality is provided by the hpocket service. As suggested by recent pocket detection methods such as Concavity (14) or

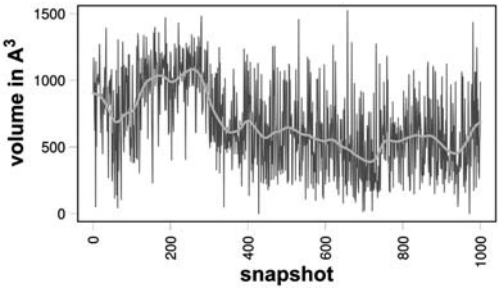


Figure 6. Volume of the P38-binding site during a 50 ns trajectory. Gray line, volume estimated for each snapshot; light gray line: smoothed volume.

Metapocket (16), taking into account the pocket surrounding residue conservation helps to refine precisely a ligand binding site. With hpocket, a complementary approach is proposed, allowing pure geometry-based pocket prediction on homologous structures.

In terms of ergonomics, the design of an online service dealing with such ensembles of structures and analysing dynamic aspects of pockets is a challenging question. The proposed suite of services included in fpocket is, as illustrated in a complex use case considering the analysis of a molecular dynamics trajectory, an efficient starting point. Interestingly, compared to isolated *default* servers, the fpocket, hpocket and mdpocket services are also provided in the Mobyle environment (33). This integration allows easy data pipelining between different applications of the Mobyle portal, including the server presented throughout this article. Presumably, such an integration allows more flexible data handling for the end users, compared to static servers, but could be enhanced. A perspective to gain in interactivity seems to lie in availability of online elementary chainable tools, a direction we are investigating.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Axel Bidon-Chanal, Ana Oliveira, Javier Luque and Xavier Barril from the University of Barcelona for contributions and constructive critics during the mdpocket conception and providing the MD trajectories, Robert Hanson (Jmol) and Mike Harsthorn (OpenAstexViewer) for their kindness and software adaptation to our request.

FUNDING

Generalitat de Catalunya (to P.S.); Conseil général du Loiret (to V.L.G.); INSERM UMR-S 973 recurrent funding (to J.M., P.T.); University Paris Diderot (RPBS servers supporting fpocket). Funding for open access charge: INSERM UMR-S 973 recurrent funding.

Conflict of interest statement. None declared.

REFERENCES

1. Levitt,D.G. and Banaszak,L.J. (1992) Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.*, **10**, 229–234.
2. Laskowski,R.A. (1995) SURFNET: a programm for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330, 307–308.
3. Liang,J., Edelsbrunner,H. and Woodward,C. (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **7**, 1884–1897.
4. Brady,G.P. Jr and Stouten,P.F. (2000) Fast prediction and visualization of protein binding pockets with pass. *J. Comput. Aided Mol. Des.*, **14**, 383–401.
5. Venkatachalam,C.M., Jiang,X., Oldfield,T. and Waldman,M. (2003) Ligandfit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.*, **21**, 289–307.
6. Nayal,M. and Honig,B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, **6**, 892–906.
7. Weisel,M., Proschak,E. and Schneider,G. (2007) Pocketpicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.*, **1**, 1–17.
8. Kim,D., Cho,C.-H., Cho,Y., Ryu,J., Bhak,J. and Kim,D.-S. (2008) Pocket extraction on proteins via the voronoi diagram of spheres. *J. Mol. Graph. Model.*, **26**, 1104–1112.
9. Schwarz,B., Albou,L.P., Wurtz,J.M., Pock,O. and Moras,D. (2008) Defining and characterizing protein surface using alpha shapes. *Proteins*, **78**, 1–12.
10. Le Guilloux,V., Schmidtke,P. and Tuffery,P. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics.*, **10**, 168.
11. Laurie,A.T.R. and Jackson,R.M. (2005) Q-SteFnder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics.*, **21**, 1908–1916.
12. An,J., Totrov,M. and Abagyan,R. (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics*, **4**, 752–761.
13. Huang,B. and Schroeder,M. (2006) LIGSITECSC: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct. Biol.*, **6**, 19.
14. Capra,J.A., Laskowski,R.A., Thornton,J.M., Singh,M. and Funkhouser,T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.
15. Skolnick,J. and Brylinski,M. (2009) Findsite: a combined evolution/structure-based approach to protein function prediction. *Brief. Bioinform.*, **10**, 378–391.
16. Huang,B. (2009) Metapocket: a meta approach to improve protein ligand binding site prediction. *OMICS*, **13**, 325–330.
17. Hessling,M., Richter,K. and Buchner,J. (2009) Dissection of the ATP-induced conformational cycle of the molecular chaperone HSP90. *Nat. Struct. Mol. Biol.*, **16**, 287–293.
18. Sherman,W., Beard,H.S. and Farid,R. (2006) Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.*, **67**, 83–84.
19. Brylinski,M. and Skolnick,J. (2008) What is the relationship between the global structures of apo and holo proteins? *Proteins*, **70**, 363–377.
20. Cuneo,M.J., Beese,L.S. and Hellinga,H.W. (2008) Ligand-induced conformational changes in a thermophilic ribose-binding protein. *BMC Struct. Biol.*, **8**, 50.
21. Eyrisch,S. and Helms,V. (2007) Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.*, **50**, 3457–3464.
22. Eyrisch,S. and Helms,V. (2009) What induces pocket openings on protein surface patches involved in protein-protein interactions? *J. Comput. Aided Mol. Des.*, **23**, 73–86.
23. Brylinski,M. and Skolnick,J. (2008) A threading-based method (findsite) for ligand-binding site prediction

and functional annotation. *Proc. Natl Acad. Sci. USA*, **105**, 129–134.
24. Dundas,J., Ouyang,Z., Tseng,J., Binkowski,A., Turpaz,Y. and Liang,J. (2006) CASTP: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, **34**, W116–W118.
25. Kalidas,Y. and Chandra,N. (2008) PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.*, **161**, 31–42.
26. Tseng,Y.Y., Dupree,C., Chen,Z.J. and Li,W.-H. (2009) SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns. *Nucleic Acids Res.*, **37**, W384–W389.
27. Tseng,Y.Y., Chen,Z.J. and Li,W.-H. (2010) FPOP: footprinting functional pockets of proteins by comparative spatial patterns. *Nucleic Acids Res.*, **38**, D288–D295.
28. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38, 27–28.
29. Biegert,A. and Söding,J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl Acad. Sci. USA*, **106**, 3770–3775.
30. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
31. Chandonia,J.-M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The astral compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
32. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
33. Néron,B., Ménager,H., Maufrais,C., Joly,N., Maupetit,J., Letort,S., Carrere,S., Tuffery,P. and Letondal,C. (2009) Mobyle: a new full web bioinformatics framework. *Bioinformatics.*, **25**, 3005–3011.
34. R Development Core Team. (2009) *R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing*. Vienna, Austria, ISBN 3-900051-07-0.
35. Faggiano,S., Abbruzzetti,S., Spyrakis,F., Grandi,E., Viappiani,C., Bruno,S., Mozzarelli,A., Cozzini,P., Astegno,A., Dominici,P. *et al.* (2009) Structural plasticity and functional implications of internal cavities in distal mutants of type 1 non-symbiotic hemoglobin Ahb1 from arabidopsis thaliana. *J. Phys. Chem. B*, **113**, 16028–16038.

3.4 Pocket Database & Applications

3.4.1 Introduction

The druggability prediction method associated with fpocket, presented in section 3.1 of this thesis, can now be used to detect all cavities on all protein structures in the PDB. This knowledge can then be analysed for identifying novel targets, until now not considered for drug-discovery.
A main objective of the work presented in this thesis was the discovery of druggable cavities situated on the surface of protein-protein complexes. Studying these particular cavities is part of a bigger ongoing research line in our group. This project is based on the following observations:

- Especially structure based drug discovery projects focus on known binding sites, like enzyme active sites or known substrate binding sites
- The power of using alternatives to classical equilibrium binding strategies is currently underestimated [Swinney and Anthony, 2011]
- *"In situations in which the dynamic actions of the drug substance stimulate, or inhibit, a biological process, it is necessary to move away from the descriptions of single proteins, receptors and so on and to view the entire signal chain as the target" [Imming et al., 2006]*

From a structure based viewpoint, addressing these issues is a challenging task. Here I propose a first work towards an integrative knowledge-base of structure information on proteins and their cavities.

Structural databases and pitfalls Structural databases of macromolecules are merely used to initially identify new targets for drug discovery purposes. This is due to several reasons. First, rational drug discovery is focused on cognate binding sites and active sites. Second, databases gathering structural, chemical and genetic information on proteins are hardly consistent with each other. Thus the development of a global knowledge-base to efficiently address target assessment prior to drug discovery projects is not straightforward. For instance, the Protein Data Bank is an incredibly rich resource and is central in such an undertaking. However, it was also until recently the major limiting factor. The lack of consistency in the structural deposition process, annotation and data organisation in the PDB make it difficult to handle this data in automatic pipelines. This can be exemplified by the observation that there are several so-called mappings that exist trying to map residues in protein structures to the actual sequence of the protein deposited in resources like Uniprot [Velankar et al., 2005, Martin, 2005, David and Yip, 2008]. A recent paper of Robert Sheridan and co-workers resumes important problems one encounters while using the PDB, especially in the study of cavities [Sheridan et al., 2010]. The following paragraph lists main issues identified in that paper, completed with further observations.

PDB caveats related to automated pocket prediction:

- The PDB is heavily biased towards some particular targets. Proteins that are easy to purify and crystallize are more likely to be in the PDB. Probably the heaviest bias is towards therapeutically interesting targets, with an obvious relation to immediate impact on pharmaceutical research.
- There is never a guarantee that what appears in the crystal structure is the species to which a drug-like compound would actually bind. Sheridan et al. cite the example of HIV integrase inhibitors [Sheridan et al., 2010] that do not only bind to the protein but to the complex of protein and DNA. Connected to this issue are incomplete structures, where loops are too flexible to be crystallized and important residues are mutated to help crystallisation.
- There are numerous inconsistencies from one PDB structure to the other, namely, residue numbering rules that are crystallographer dependent (several residues can have the same residue number), atom ordering inside a residue and atom naming.
- Several issues persist regarding treatment of what is not a protein or a macromolecule (DNA/RNA). Ligands, including drugs, co-factors or ions are all treated as heterogroups. Furthermore, with recent corrections in the PDB, several ligands were split into independent residues (fragments), making thorough identification more difficult. There is no indication to distinguish co-factors from drugs. On the contrary, small peptides that act as ligands are considered to be proteins.
- Proteins and binding sites are flexible. Using automatic pocket detection methods can yield a single pocket in one structure and split pockets on other structures.
- Crystal structures are deposited without explicit indication whether a protein complex is *(i)* constitutive or transitory, *(ii)* biologically relevant complex or due to crystal packing.
- The previous caveat also makes it more difficult to be able to consider only one chain in a protein complex and perform pocket detection to identify putative PPI cavities.

Linking different databases Ideally, structural information should be easily linked to other data sources. For example what are the phenotypic implications of altering a given protein in a disease-related pathway? This can currently be accomplished by combining several inconsistent resources with each other.
Another central issue for structure based drug discovery is the linkage of chemical space to actual targets. While resources like DrugBank [Wishart et al., 2006], BindingDB [Liu et al., 2007], PDBBind [Wang et al., 2005], Relibase [Hendlich et al., 2003] or BindingMOAD [Hu et al., 2005] have their importance in analysing known drug binding sites, they have only limited significance in

the discovery of novel putative pockets.
Among the best known examples of a comprehensive pocket database is the Relibase, Cavbase and Waterbase trio, published and maintained by the group of Gerhard Klebe. More recently two commercial main players in computational drug discovery, CCG and Schrödinger, have made first attempts to provide comprehensive databases focussing on binding sites found in the PDB. While CCG has a product called PSILO (`http://www.chemcomp.com/psilo-info.htm`), Schrödinger announced intentions to strengthen efforts in the field. This shows that there is an interest in the pharmaceutical industry for this kind of integrated databases. PSILO includes very interesting ideas allowing browsing and queries on ligand and 3D structural data, similar to Relibase. Furthermore pocket similarity searches are implemented and users can annotate data in a collaborative manner. Despite the existence of PSILO, Relibase and Cavbase, none of them is available free of charge. Other notable ensembles focusing on the analysis of protein binding sites alone are EBIeMotif [Golovin and Henrick, 2008] and CREDO [Schreyer and Blundell, 2009].
In an interesting approach, Panjkovich and Daura show the usefulness of such integrative structural databases [Panjkovich and Daura, 2010]. Through analysis of known binding sites in different protein families, the authors derived a measure of pocket conservation within each protein family combining structural and sequence information. This measure was then subsequently used to detect conserved putative allosteric binding sites within these families. Here a related aim is followed using slightly different approaches.

3.4.2 Objectives

Construct a relational database on automatically detected pockets The primary objective of this work is to use an automatic pocket identification algorithm (fpocket) to systematically identify cavities on all PDB structures. The resulting cavities and corresponding properties should be arranged in a relational database to allow easy and fast querying of the data. Connected to information on the cavities, extensive information from the PDB should be added to enhance applications of the database.

Incorporate sequence information Uniprot should be used to map known protein sequences to the actual PDB structures. This sequence information can then later be used to select similar (Uniref 50, 90 or 100) proteins and map binding sites to the sequence to see their relative conservation among different structures.

Incorporate quartenary structure assessment A current major limitation of structures deposited in the PDB is the fact that hetero-multimeric or homo-multimeric complexes can be artefacts and not biologically relevant. The biological relevance can be assessed via the Biounit entry in each PDB file or via automated prediction tools, which have to be included into the database.

Suitability for punctual intuitive navigation A relational database is usually difficult to exploit for non-initiated people with the entity-association scheme of the database. Particular emphasis should be put on the development of an interface to the database that allows easy navigation in the cavity space of the PDB, visualisation of 3D structures, compounds and pocket properties.

3.4.3 Results

Description of the data

The PDB version of winter 2008 was used to run fpocket systematically using default cavity detection parameters and the -d flag producing pre-formatted output for database creations. Note that fpocket performs some *intelligent* cleaning operations on PDB structures already. Among them, fpocket strips water molecules and non standard residues with the HETATM designation in the PDB file, unless they are defined in the structural cofactor list included in fpocket. For more information see [Le Guilloux et al., 2009]. Using this -d flag fpocket produces two outputs for each protein. One line per pocket describing the cavity using PDB Ids, cavity Ids and all scores and descriptors computed by fpocket. The other output is given in form of structural files (PDB and PQR) representing the 3D coordinates and radii of alpha spheres as well as identified pocket atoms and pocket environment atoms. For the sake of maintainability this 3D structural information was not stored inside the relational database itself, but in a folder structure on the hard drive inspired by the PDB layout. For each pocket a unique identifier is defined, hereafter called pocket serial and stored to a PostGreSQL database. This pocket serial serves as primary key to gain quick, direct and unambiguous access to the pocket descriptors and further information. For a detailed entity-relationship diagram refer to figure ea.png available on the CD. This figure is not available in the print version of the thesis due to its dimensions. Main properties of this architecture will be resumed within the text.
Cavity information is stored in a table named *pdb_d_cavities*. Associated also to each pocket is also a PDB Id and a cavity Id (from fpocket). Using the PDB Id, *pdb_d_cavities* is linked to the table *pdb* containing several details on a given PDB structure, like experimental methods used, resolution of the structure, title and headers as well as digital object identifiers for the primary publication of the structure.
The pocket database distinguishes two types of cavities. One cavity can contact a single chain of a protein or several chains of the same or different proteins. In case a cavity contacts a single chain an entry in the table *pdb_d_cavities_single_chain* is used to gather information on which chain the pocket lies (chain identifier from the PDB file), what type of chain it is (protein, DNA or other) and how long the chain is in total (number of residues). In case the pocket contacts two or more chains, an entry in table *pdb_d_cavities_inter_chain* is written. Last, if a small molecule is found in the pocket and if it is not a known cofactor, then its residue name is tracked in the table *pdb_d_cavities_lig_het_tag*.
In theory, each residue occurring as HETATM in a PDB file is listed in the table *pdb_het*. Here, several inconsistencies have been observed between the

PDB and the dictionary of HETATM residue names provided by the PDB. Ideally relational constraints should be applied to this table, but as the aim is to include as much data as possible, the data is loosely added and software filters are used to treat data inconsistencies rather than the rigid database relational constraints. The *pdb_het* gathers not only data on the residue name, but also SMILES strings of small molecules and molecular properties calculated using RDKit[Landrum, 2011], like counts of various atoms (heavy atoms, polar atoms), ring counts, rotatable bond counts and the Crippen LogP.
An important component of the database is built from Uniprot data. Uniprot was downloaded and a parser written. Next to references between Uniprot Ids, Uniprot Accessions and Uniref Accessions, sequence information is stored into the database, as well as various annotations, like known active or binding sites, mutants etc.
Furthermore, a part of Kegg was included into the pocket database [Kanehisa et al., 2010]. Here signalling pathways, drug-protein interactions and protein-disease relations were integrated.
To be able to distinguish biologically relevant quarternary structures from experimental artefacts the *pisa_interface* table contains only chain partners in proteins that are predicted to be relevant using PDBePISA [Krissinel and Henrick, 2007]. Furthermore, various uniprot families and assigned enzyme classes are stored. Last, most importantly, a PDB to Uniprot mapping was included. This mapping allows to match the PDB sequence to the sequence in the Uniprot database. In this work, the mapping provided by Andrew Martin was used [Martin, 2005].
Finally, the residues of automatically identified cavities that are in contact with alpha spheres are gathered together in a table names *pdb_d_cavities_resnum_chain*. This tracking allows very quick mapping of cavities to the primary amino-acid sequence.
Due to inconsistencies between all these databases (doubled entries, renamed identifiers mostly in provided flat files) all components, i.e. the PDB, Kegg and Uniprot are separated relationally in the entity relation scheme of the database.

Intuitive navigation

Various ways of how to exploit this resource can be considered. Here the database was developed as in-house tool to mine cavities in the whole PDB. A PyQt graphical user interface [Riverbank, 2010] was developed to efficiently and intuitively browse through this data. On figure 3.21 a screen-shot of the Graphical User Interface (GUI) is shown. The main interface gathers different information together into one window. Window components are numbered in figure 3.21 and explained here:

1. Search field: Here the user can search either for a PDB code or for a keyword used to describe PDB structures in the Protein Data Bank itself.

2. 2D Ligand viewer: If a small molecule is found inside a binding pocket that has been selected by the user, the 2D structure of the ligand is shown here.

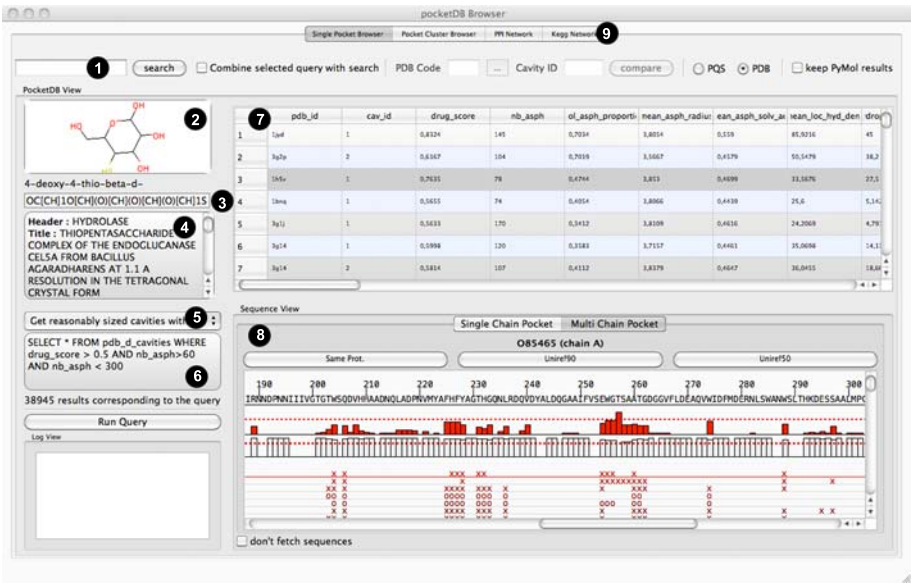


Figure 3.21: Graphical user interface for the pocket database

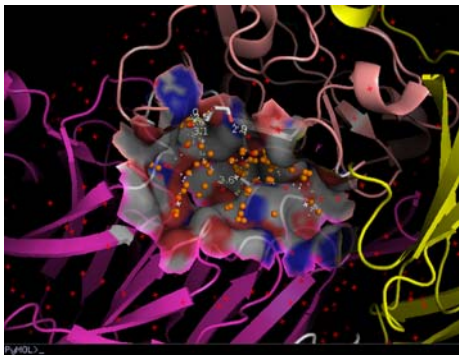


Figure 3.22: Cavity of interest is viewable in independent PyMOL session

3. Molecule name and SMILES: If a small molecule is found inside a binding pocket that has been selected by the user, here the name of that molecule and the corresponding SMILES string is shown.

4. Resumé of the structure: Here information on the title / header of the PDB file, a link to the publication and information from uniprot is shown. This section gives a first insight on what the protein is, how the structure was determined and if there are annotated functions and binding sites.

5. Predefined queries: To facilitate the use of common searches, a set of pre-defined queries have been implemented and are available from the combo-box.

6. SQL input field: If the user is familiar with the database design, then he

can write direct SQL code and query the database (by pressing the "Run Query" Button). If the user selects a predefined query from field 5, then this query is also shown in this field.

7. Results panel: Here results from the SQL query are shown in a tabular formatting. This table is interactive and if a pocket is shown per line (depending on the query), then a click on that pocket opens several subsequent details, like a 2D structural view of the ligand in (2), a 3D structural view of the protein and the pocket in PyMOL [DeLano, 2002], a sequence view in (8) and if available other details in (9), which will be explained hereafter.

8. Sequence panel: If the user selects a pocket from the results panel, then the GUI opens the protein sequence (from Uniprot). Onto that sequence all pockets from all structures for that same protein are then mapped. Each line in this panel corresponds to a distinct cavity, ordered by drug score. A "x" is drawn for each amino acid in contact with a cavity containing a ligand and a "o" if there is no ligand inside the pocket. The sequence panel is also interactive, allowing the user to select a pocket by clicking and thus viewing details and the 3D structure of it.

9. Supplementary information tabs: Several tabs are shown here and the tab activated by default is the window shown in figure 3.21. A second tab called *"Pocket Cluster Browser"* is explained in further detail in the next subsection, as well as the PPI Network tab. Figure 3.22 shows an example of a cavity opened with the pocket browser. The protein is generally represented in cartoon and for the surface portion in contact with the cavity (defined by alpha sphere centres) the solvent accessible surface is shown.

Pocket Clustering

As previously shown the GUI contains a tab named *Pocket Cluster Browser*. Indeed in the default view, all pockets are shown for all proteins on the same protein (or PDB structure) providing an important amount of information especially when a protein has several distinct cavities on its surface. If one wants to focus on a particular cavity, for instance a known binding site and know all PDB structures containing ligands, then this can be easily spotted in the Sequence Viewer for instance.
Information in the PDB and also in the pocket database presented here is very redundant. Several times the same or a very similar cavity is found on either different homo-multimeres on the same structure or on different structures of the same protein. Thus mining this redundant data for extracting statistics on pocket properties is not straightforward. To facilitate the identification of redundant pockets in the database, a protocol has been implemented that allows clustering of redundant cavities to so called pocket clusters. The *Pocket Cluster Browser* as part of the GUI, simply allows to visualize these clusters.

Creating pocket clusters As previously seen, each cavity in the pocket database is mapped to the sequence of the protein it is situated on. For the sake of simplicity only cavities situated on a single protein chain are considered first. Cavities in the pocket database are in contact with nearly 11 residues on average. For clustering, only cavities in contact with more than 4 residues are considered to avoid using too small pockets. Using the pocket database it is straightforward to gather all cavities on a particular protein. Now one can assess how much one pocket overlaps with another once it has been mapped to its primary protein sequence. If two pockets significantly overlap with each other, then they are clustered together. Using an agglomerative clustering approach, finally separate pocket clusters can be identified defining single non redundant zones on the primary sequence corresponding to a cavity. As this cavity is seen on several structures of the same protein it can be valuable to analyse these several occurrences. The *Pocket Cluster Browser* is shown in figure 3.23. The window is updated with information if the user selects

Figure 3.23: Graphical user interface for the pocket database cluster browser

a particular cavity in the database table from initial results in the default window of the navigator. Several aspects can be analysed using the *Pocket Cluster Browser* (numbering is referring to figure 3.23):

1. Query View: Here the SQL query is shown corresponding to selecting all cavities in the same cluster as the selected cavity by the user.

2. Database Results View: In analogy to the default viewer, here pocket details for all cavities of the same cluster are shown.

3. Property Analyzer: This is a graphical tool included here to analyse the distribution of various properties of cavities within the same cluster. For

instance, if one intends to analyse the druggability (as shown in this example), a histogram of druggability values of all cavities in the cluster is displayed.

4. Sequence Viewer: Here again the cavities are mapped to the sequence as already shown on figure 3.21. However, this time only the cavities in the same cluster (so the same pockets on different structures) are shown. This allows very quick assessments of the correctness of pocket clustering using the sequence mapping.

5. Ligand Viewer: Here all ligands found in pockets constituting the pocket cluster are shown using a 2D representation. It should be noted that these ligands are interactive and that a click on them loads the 3D structure of the pocket and the protein onto the current pocket, allowing this way straightforward pocket alignments and superpositioning of ligands inside binding sites using PyMOL.

Protein interaction networks and diseases

Another factor included in the pocket database is information on disease relation of proteins and known interaction pathways between different proteins. Here again with a click on a specific pocket in the default window of the navigator, known protein protein interaction networks are loaded into the *PPI Network* tab. Figure 3.24 shows an example from a selection of the ATP

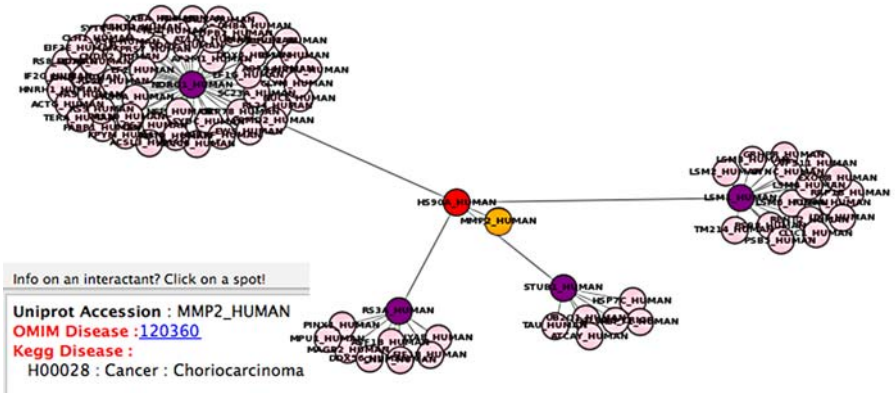


Figure 3.24: Example of a protein protein interaction network centered on HSP90 (red node).

binding site of Heat schock protein 90. Each node in that representation corresponds to a protein labeled with the Uniprot Id. The selected protein itself is coloured in red in the protein network viewer. Next, every level of subsequent interactions is coloured differently. By default 2 levels of subsequent interactions are fetched, but the spread of the interaction network to extract can be user defined. The protein interaction map is interactive and a click on a given

node fetches disease relation from the Online Mendelian Inheritance in Man compendium [Hamosh et al., 2005] and the Kegg disease database.

Applications

While the pocket database and the graphical user interface allow punctual easy browsing and visualisation of cavities of a protein it can also be used to systematically sieve through all pockets and identify pockets and proteins with a certain set of characteristics.
An ongoing work intends to estimate the druggability of different types of pockets and refine the druggable genome, initially defined by Hopkins and Groom [Hopkins and Groom, 2002]. This can be done via statistical analysis of binding sites in the pocket database. From such an analysis it was found that around 20% of all proteins in the PDB contain at least one druggable cavity situated on a single peptidic chain. This estimate is relatively close to the 14% proposed by Hopkins and Groom, considering that the current estimate of the size of the human genome is around 22.000 genes [Hesman Saey, 2010]. Further analysis shows that there is a 10% probability to find cavities on protein interface to either inhibit or stabilise the interaction. The stabilisation of such an interaction is of particular interest to our group and is analysed in more detail in the following paragraph.

Identify PPI glue sites: As part of a major ongoing project in our research group, the pocket database and the previously developed druggability score associated with each pocket can be used to identify special cavities. These pockets are situated on the surface of a protein-protein complex encompassing both proteins. The main aim of this research project is to target these pockets to fit a small drug-like molecule and subsequently stabilise the protein-protein interaction. Stabilising protein-protein interactions can be understood as of making them stickier, thus the accronym PPIglue is used to denominate these particular binding sites.
Untill today very few cases are known, where a small molecule triggers a stabilisation of a protein-protein complex via direct interaction. Pommier and Cherfils resumed a set of systems where natural products are used to trap protein-protein complexes in transition states so that they are unable to complete their biological function [Pommier and Cherfils, 2005]. Despite the existence of these examples occurring naturally, little interest has been shown by pharmaceutical industry in exploiting such transient pockets for drug-discovery. Knowing that a significant amount of medicines are using kinetic or conformational trapping mechanisms [Swinney and Anthony, 2011], protein-protein complex stabilisation appears as a viable alternative approach in modern rational drug discovery. The pocket database presented in this section of the thesis serves primarily the purpose of discovery of such cavities. The integration of all different datasources into one single framework allows us to query for such cavities with specific properties.
To perform this research several assumptions and approximations have been made. First only reasonably sized (number of alpha spheres between 100 and

400) cavities have been considered for this study. Next, only cavities with a drug score higher than 0.5 have been deemed druggable. Last, the cavity was allowed to have a maximum flexibility score [Le Guilloux et al., 2009] of 0.2, reducing the probability of sieving through very flexible pockets. Furthermore, only hetero-multimeric interfaces have been considered using Uniref 50 as reference sequence for comparison of the two protein chains. Last, the hetero-dimer / multimer had to be non-obligate, meaning that both proteins can exist alone and were crystallized separately.
Prior to the final analysis of results, several types of systems were excluded as irrelevant, notably immune system related antibodies, membrane proteins, proteins related to blood clotting and toxins. This protocol resulted in a list of 248 redundant pockets on 43 different systems among which 4 were discarded as false positives (constitutive dimers with wrong Uniprot assignments) resumed hereafter in table 3.6. Interestingly, some of the known natural com-

No.	PDB Code	Comments
1	2c37	Sulfolobus solfataricus: Complex between Exosome Exonuclease 1 and 2, part of multimeric complex
2	1k7e	Salmonella typhimurium: Tryptophan synthase alpha & beta chain interface co-crystallized with inhibitor
3	2pnz	Pyrococcus abyssi: Complex between Exosome Exonuclease 1 and 2, similar to No. 1
4	2q97	Toxoplasma gondii: Toxofilin rabbit actin complexe, ATP binding site
5	1f2r	Mus muculus: Caspase activated DNAse complexed with protein inhibitor
6	1rj9	Thermus aquaticus: Complex between Signal recognition protein and Signal recognition particle protein, GTP binding site
7	1lb1	Homo sapiens: Complex between RhoA and DBS, known GTP binding site
8	2f9d	Homo sapiens: Complex between Splicing factor 3B & Pre-mRNA branch site protein p14, binding site involving 4 protein chains
9	1cs4	Mammals: Heterodimer of adenylate cyclase, known forskolin binding site [Tesmer, 1997]
10	2g81	Bos taurus: Trypsin with trypsin inhibitor, hexaethylene glycol binding in the interface stabilizes the complex [Barbosa et al., 2007]
11	1lw6	Bacillus amyloliquefaciens: Subtilisin BPN' with Chemotrypsin inhibitor
12	2fcw	Homo sapiens: Complex between LDL and RAP. MPD molecule found in interfacial cavity
13	2omi	Homo sapiens: Insulin multi heteromeric complex cavity
14	1o9a	Homo sapiens: Fibronectin - fibronectin binding protein complex
15	1gua	Homo sapiens: RAP C-RAF1 complex. GTP binding site
16	2k8f	Homo sapiens: p53 in complex with histone acetyltransferase p300
17	3ci5	Dictyoselium discoideum: Actin gelsolin complex
18	2arp	Homo sapiens: Activin A, follistatin complex
19	1a0o	Escherichia coli: CheA/ CheY complex
20	2za4	Bacillus amyloliquefaciens: Barnase Barstar complex
21	1efv	Homo sapiens: Electron transfer flavoprotein, AMP binding site is part of interfacial cavity
22	1ira	Homo sapiens: Interleukin 1 Receptor & antagonist complex
23	1pvh	Homo sapiens: Interleukin 6 Receptor & leukemia inhibitory factor
24	2fin	Homo sapiens & Rabbitpox virus: cytokine A4 chemokine inhibitor complex
25	2e30	Homo sapiens: Calcium binding protein p22 & Sodium hydrogen exchanger complex
26	1h25	Homo sapiens: CDK2 Cyclin A complex
27	2h4y	Homo sapiens: Caspase 1, p20 & p10 complex, active site found with inhibitor
28	3c7u	E.coli & Streptomyces clavuligerus: Beta lactamase & beta-lactamase inhibitory protein, binding site close to known pocket on Beta lactamase, but occupied by BLIP
29	1oxb	Saccharomyces cerevisiae: YPD1/SLN1, pocket occupied by SO_4
30	3fap	Homo sapiens: FKBP12-rapamycin-FRB ternary complex, known PPI stabilizer rapamycin
31	3c0r	Homo sapiens & Saccharomyces cerevisiae: Ubiquitin & Ubiquitin thioesterase OTU1 complex
32	2fuh	Homo sapiens & Rattus norvegicus: Ubiquitin & Ubiquitin-conjugating enzyme E2 D3
33	2f4m	Mus muculus: peptide N-glycanase-HR23 complex
34	1bi8	Homo sapiens: CDK6-CDK inhibitor complex. Pocket encompassing ATP binding site
35	3h9r	Homo sapiens: Activin receptor 1 - FKBP1A complex, empty pocket coincides with rapamycin binding site known in FKBP
36	1wr6	Homo sapiens & Bos taurus: Ubiquitin GGA3 complex, pocket on the interface of 3 chains
37	1s9d	Homo sapiens & Bos taurus: Arno & ADP-Ribosylation Factor 1. Pocket binds brefeldin A, a known PPI stabilizer
38	1l3e	Homo sapiens: Hypoxia Inducible Factor 1 - p300

Table 3.6: Protein-protein interfacial druggable pockets identified via systematic screen of the pocket database.

pound pockets discussed in [Pommier and Cherfils, 2005] are found via this protocol in the pocket database, validating the protocol used here. Using our protocol on the pocket database, we are able to identify 3 different systems already described in related literature. Namely, brefeldin A, rapamycin and forskolin binding sites [Pommier and Cherfils, 2005] have been identified alongside with several other pockets that either did not have known co-crystallized ligands in the interfacial cavity or other molecules (GTP, AMP etc...). The fact that independently of any information on ligands the protocol is able to identify these known PPI stabilizers is a validation for the detection protocol, the pocket database integrity and the druggability assessment.
In a subsequent work we manually selected 3 systems out of the series of complexes shown in table 3.6. Given our restricted possibilities, not all possible systems are currently investigated.
These three examples, namely the Rap Raf complex (15), the CheA/CheY complex (19) and the CDK6/CDK6 inhibitor complex (37), are investigated with a more generic approach for druggability assessment called MDMix [Seco et al., 2009]. Using molecular dynamics of the protein complex several solvent mixtures are considered and the relative occupancy of each solvent type and molecule is assessed. This occupancy is then translated to a maximum affinity that a drug-like molecule could have for a binding site. Furthermore, interaction hotspots for various interaction types (hydrogen bond acceptors / donors, hydrophobic groups etc..) can be pointed out by the method.

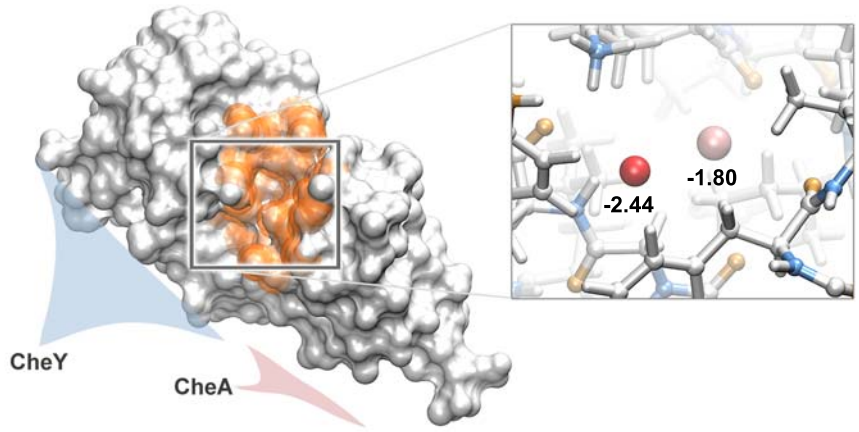


Figure 3.25: Overview of the CheA CheY complex in E.coli. Detailed view of the cavity found on the interface with 2 strong interaction hotspots (spheres) for H-Bond acceptors (-2.44 kcal/mol) and hydrophobic groups (-1.80 kcal/mol)

Chea/Chey - proof of principle As starting point we decided to use the CheA/CheY complex as reference case to proof that PPI stabilisation can be rationalised and such sites found via rational protocols. This choice has several reasons. First, the organism E.coli is well known and easy to manipulate

experimentally. Both proteins are easy to produce and the complex formation and stabilization can be assessed with techniques like Biacore, but also possibly via macroscopic in-vivo essays (motility essays).
On this protein-protein complex (PDB structure 1a0o) MDMix was run within our research group and 2 interaction hotspots in the bottom of the cavity were identified. Figure 3.25 A shows the location of the binding site on the complex. The two most important hotspots identified are shown on figure 3.25 B. These two locations were used as pharmacophoric restraints for a systematic docking of all ligands from the vendor SPECS in our in-house small molecule database CDBM. The docking was performed using both rDock [Morley and Afshar, 2004] and Glide(SP) [Halgren et al., 2004]. Currently, identified ligands via this virtual screen are assessed experimentally.

Novelty Binding site databases have existed for a long time and are reported in litterature. However, only few works encompass the whole pocketome, meaning all pockets on all proteins, not only cognate binding sites. Here, a unique resource is presented, associating data from different renowned data-sources to a loose ensemble. This data can be used to detect and characterise yet uncharacterised pockets.

Limitations The consistency of the database is difficult to achieve because of the inconsistency between different data-sources. This leads sometimes to entries in the PDB with missing corresponding Uniprot information, for example. Structural data corresponds to data from late 2008. An update is required but a thorough validation protocol for data-integrity after such an update has yet to be developed. The database is not distributed in its current form.

Bibliography

João Alexandre R G Barbosa, Luciano P Silva, Rozeni C L Teles, Gisele F Esteves, Ricardo B Azevedo, Manuel M Ventura, and Sonia M de Freitas. Crystal structure of the Bowman-Birk Inhibitor from Vigna unguiculata seeds in complex with beta-trypsin at 1.55 A resolution and its structural properties in association with proteinases. *Biophysical journal*, 92(5):1638–50, March 2007.

Fabrice P A David and Yum L Yip. SSMap: a new UniProt-PDB mapping resource for the curation of structural-related information in the UniProt/Swiss-Prot Knowledgebase. *BMC bioinformatics*, 9:391, January 2008.

Warren L DeLano. The PyMOL Molecular Graphics System, 2002.

Adel Golovin and Kim Henrick. MSDmotif: exploring protein sites and motifs. *BMC bioinformatics*, 9(1):312, January 2008.

Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–9, March 2004.

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue): D514–D517, 2005.

Manfred Hendlich, Andreas Bergner, Judith Günther, and Gerhard Klebe. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *Journal of molecular biology*, 326(2):607–20, February 2003.

Tina Hesman Saey. More Than A Chicken, Fewer Than A Grape - Science News, 2010.

Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature reviews. Drug discovery*, 1(9):727–30, September 2002.

Liegi Hu, Mark L Benson, Richard D Smith, Michael G Lerner, and Heather A Carlson. Binding MOAD (Mother Of All Databases). *Proteins*, 60(3):333–40, August 2005.

Peter Imming, Christian Sinning, and Achim Meyer. Drugs, their targets and the nature and number of drug targets. *Nature reviews. Drug discovery*, 5(10):821–34, October 2006.

Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(Database issue):D355–60, January 2010.

Evgeny Krissinel and Kim Henrick. Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3):774–97, September 2007.

Greg Landrum. RDKit : A software suite for cheminformatics , computational chemistry , and predictive modeling. *Components*, 2011.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10:168, January 2009.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic acids research*, 35(Database issue):D198–201, January 2007.

Andrew C R Martin. Mapping PDB chains to UniProtKB entries. *Bioinformatics (Oxford, England)*, 21(23):4297–301, December 2005.

S David Morley and Mohammad Afshar. Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock. *Journal of computer-aided molecular design*, 18(3):189–208, March 2004.

Alejandro Panjkovich and Xavier Daura. Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC structural biology*, 10:9, January 2010.

Yves Pommier and Jacqueline Cherfils. Interfacial inhibition of macromolecular interactions: nature's paradigm for drug discovery. *Trends in pharmacological sciences*, 26(3):138–45, March 2005.

Riverbank. PyQt, 2010.

Adrian Schreyer and Tom Blundell. CREDO: a protein-ligand interaction database for drug discovery. *Chemical biology & drug design*, 73(2):157–67, February 2009.

Jesus Seco, F Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *Journal of medicinal chemistry*, 52(8):2363–71, April 2009.

Robert P Sheridan, Vladimir N Maiorov, M Katharine Holloway, Wendy D Cornell, and Ying-Duo Gao. Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *Journal of chemical information and modeling*, 50(11): 2029–40, November 2010.

David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews. Drug discovery*, 10(7):507–19, January 2011.

J. J. Tesmer. Crystal Structure of the Catalytic Domains of Adenylyl Cyclase in a Complex with Gs·GTPS. *Science*, 278(5345):1907–1916, December 1997.

S Velankar, P McNeil, V Mittard-Runte, A Suarez, D Barrell, R Apweiler, and K Henrick. E-MSD: an integrated data resource for bioinformatics. *Nucleic acids research*, 33(Database issue):D262–5, January 2005.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The PDBbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–9, June 2005.

David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34 (Database issue):D668–72, January 2006.

3.5 The Pocketome

3.5.1 Introduction

This last section of my thesis presents ongoing work for 4 years. Theoretical basics were settled during my short stay at Sanofi Aventis and the following years allowed me to complete the project to its current state. Nearly all previously developed tools and databases come to use to provide a unique resource, *the pocketome*. Although this work has not been published in paper format yet, I deemed it sufficiently advanced to report first validation and results in this thesis. Furthermore, it has already been presented at the Gordon Research Conference on Computer Aided Drug Design 2011 in Vermont-USA.

The dominant paradigm in drug discovery is the concept of designing maximally selective ligands to act on individual drug targets [Hopkins, 2008]. This paradigm directly implies that a small molecule acts only on one target and solely induces an effect there. However, the cell is a crowded environment where proteins encounter all types of other macro-molecules, solvent, other solutes and small molecules. Necessarily, drugs bind transiently to other macromolecules. Such usually unwanted interactions are in most cases weak and don't induce any effect on the transient interaction partner. However in some cases they can also cause severe side-effects or toxicity.
Another major issue in current drug development processes is the efficiency of small molecules [Kola and Landis, 2004]. A very specific drug might be able to bind to a single target and alter the function of it. However, systems biology approaches have shown that the effect produced by altering one protein can be cancelled by the robustness of biological interaction networks [Albert et al., 2000, Kitano, 2007, Maslov and Ispolatov, 2007].
These observations gave rise to two research areas, polypharmacology and chemogenomics [Kola and Landis, 2004, Bender et al., 2007].

Chemogenomics is a discipline promoting systematically screens of series of ligands against multiple targets, usually of the same family. This aims to generate activity profiles of each compound versus multiple proteins and select those with optimal selectivity. Such studies allow to relate targets via ligand activity and similarity.

Polypharmacology and multi-targeting approaches intend to develop small molecules specifically targeting multiple targets either using one molecule [Bongarzone et al., 2010] or a combination of drugs to improve the efficiency of current treatments [Hopkins, 2008].

Both disciplines are intimately linked to problems outlined previously regarding toxicity and efficacy of drugs. More recently, structural genomics projects allow to foster systematically 3D structures of macromolecules involved in the functioning of a given organism. This enables us to gain exciting new insights into the complete structural space of a cell. Once all structural information is

available, also structure-based computational tools can be used to investigate if a ligand might bind to a battery of targets.
Despite the fact that the currently available structural space is incomplete and highly biased, several attempts are made to use computational tools already available to account for cross-reactivity of ligands:

- Systematic molecular docking is performed of a set of ligands on a variety of target proteins
- Ligand-protein interaction maps covering known interaction and structural space are established and enlarged via ligand similarity measures
- Pockets are compared between each other to point out structural similarities.

The study presented here will focus on the last two aspects, but especially the last one, pocket comparison.
In several aspects the project presented here has parallels to the TB-drugome [Kinnings et al., 2010]. Kinnings and colleagues created a protein-drug interaction network for *Mycobacterium tuberculosis*, M.tb. If a drug is known to bind it is added to the network connected to the corresponding protein node. Furthermore, drugs are associated to proteins if the pockets in which they are found (not necessarily in M.tb) are similar to a pocket on a protein of M.tb. Pocket similarity was calculated using SMAP [Xie et al., 2009]. On this precise organism, the drug-protein network was found to be highly interconnected and applications for drug repurposing and polypharmacology are discussed.
In the work presented here a pocket centric approach was chosen and a ligand is linked to a pocket if it is observed inside that pocket. Next a pocket should be linked to another pocket if both are similar using solely physico-chemical properties, as these are responsible for physical binding of a small molecule. To produce such a pocket-ligand interaction map, the pocketome, for all pockets currently in the PDB new approaches had to be developed.

3.5.2 Objectives

This work intends to primarily relate pockets from automatic cavity identification protocols to each other and to the ligands that are eventually co-crystallized in such pockets.

Consider non occupied pockets: A major part of chemo-genomic studies focused on known binding sites that have either been co-crystallized with a ligand molecule or that have been described previously. The primary objective of this work is to relate all pockets found with an automatic pocket detection algorithm.

PDB wide: The study has to be conducted on the whole PDB. Considering the amount of data to treat, underlying methods that have to be developed, as well as data-structures have to be optimized for large scale applications.

Use of an abstract cavity representation: Recent developments in the area of structural binding site comparison resulted in publication of various methods. Several of them consider a very restrictive representation of the cavity, where sensitivity to side chain motions can be expected. The pocket comparison method used herein should represent the cavity as a fuzzy, or abstract ensemble of interaction points of a given shape, without considering explicit side-chain arrangements and strict topology.

Open research: Similar to previous work presented in this thesis, the results from these implementations and analyses should be open to the public.

Intuitive navigation and visualisation: Given the complexity of underlying data, the way the data is presented to the user is of utmost importance. An intuitive and visual way to navigate the pocket and ligand space has to be considered for publishing results while guaranteeing free access to all resources.

3.5.3 Methods

A central part for the work presented here is the database of putative binding sites identified using the fpocket. This database is presented in more detail in secton 3.4 of this thesis. Cavities in this repository are compared to each other using a novel pocket comparison approach.

Pocket Comparison

The structure-based cavity comparison method to be developed here should address 5 main issues :

- define the pocket in a rotational and translational independent manner
- define a scoring metric for mutual overlap or sub-pocket identification
- be relatively insensitive to side-chain flipping events in binding sites
- be reasonably fast to allow high-throughput pocket comparison

Pocket Feature Detection: Prior to undergoing cavity comparison, each pocket is thoroughly analysed. During this step important physico-chemical features on the pocket surface are identified. These features, sometimes also called pseudo-centres, correspond to putative interaction points on the pocket surface. They are fetched from a predefined dictionary for each amino-acid and are detected on the automatically identified binding site. Here feature categories as implemented in FuzCav, proposed by Weill & Rognan, are used [Weill and Rognan, 2010]. For instance, a tyrosine can have an aromatic ring inside the pocket, but also the hydroxyl and the backbone carbonyle can be accessible. Note that in contrary to FuzCav, here backbone features are considered. Only features with an accessible surface area contribution to the pocket are counted. Thus for this sample tryosine an aromatic group, an hydrogen bond donor and a hydrogen bond acceptor are accessible. This process is exemplified

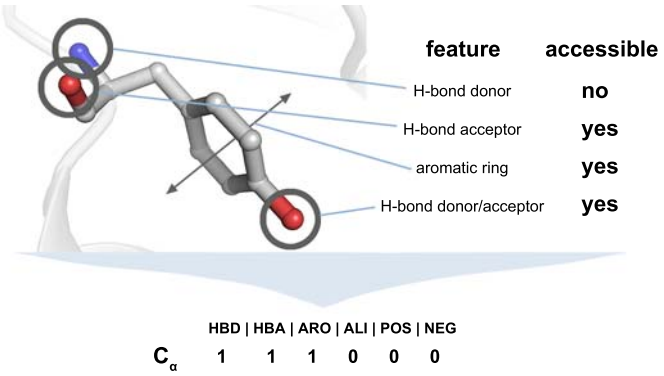


Figure 3.26: Schematic representation of the feature detection and definition process on a tyrosine. Physico-chemical features on residues lining the binding site are identified and it is assessed if they contribute to the accessible surface area of the pocket. The C_α position is tracked and the presence and absence of a feature is tracked in a 6 bit binary vector, each bit representing one possible feature.

on figure 3.26. Next, a very important level of uncertainty is introduced in the approach marking first differences with some algorithms in the field. This has been inspired by several other methods published [Xie et al., 2009, Feldman and Labute, 2010, Weill and Rognan, 2010]. For each amino-acid, only the position of the alpha carbon is retained. Thus the coordinates of all atoms on a residue are reduced to one position of the C_α. The previously identified features are known to be solvent exposed and the residue is likely to be in a concave environment. Thus, the possible arrangements of side chain atoms in such a packed space is limited, allowing the reductionist view that *"alpha carbons are enough"* [Feldman and Labute, 2010]. This level of uncertainty and reduction in positional information allows also to render the method insensitive to side-chain motions [Xie et al., 2009].
To each C_α of residues found in the pocket, a 6 bit binary fingerprint F_x is associated. Each bit corresponds to a specific type of feature that is possibly found in the pocket, hydrogen bond donors, acceptors, alipathic side-chains, aromatic groups, negatively charged and positively charged interaction points. If one or more acceptors are found on one residue, for example, the fingerprint would still only be set to 1 on the acceptor bit. All residue fingerprints are stored together with the alpha carbon positions in a fingerprint list F_l. Such a fingerprint list represents all residues in a precise pocket.

Pocket transformation to a *RIPr*: In order to compare two binding sites that are situated in two different places in 3D space, they have to be superimposed or transformed to be comparable. While this is trivial when the pockets are very similar in the same protein, it is very difficult for pockets that are not on the same protein or proteins sharing high sequence similarity. This problem is shown to be NP-hard, and even NP-complete (depending on the implementation), meaning that no efficient algorithmic solution exists to solve

that problem. Thus one has to use approximations to fulfil this task [Boukhris et al., 2009, Reisen et al., 2010]. Most of the existing methods are based on clique detection or geometric hashing techniques to superimpose two similar binding sites [Kellenberger et al., 2008]. Another approach consists in using fingerprints, also employed by techniques like FuzCav [Weill and Rognan, 2010]. In this method, pockets are reduced to a particular representation, similar to a fingerprint. One cavity has previously been reduced to a set of C_α positions in euclidean space and a list of feature fingerprints, F_l corresponding to all residues in a pocket. This list of feature fingerprints contains a limited number of possible combinations, such as that a feature fingerprint F_2 in list F_l can be the same as another fingerprint F_{34} for example. Thus a non redundant list of F_x contained in F_l is defined as F_i. Next a three dimensional space G_i is constructed having the dimension F_i, F_i, D, where D is a discrete set of distance intervals. There are 13 distance intervals ranging from 2.5 to 15.0 $\mathring{A}$ by 1.0$\mathring{A}$ steps.
All feature to feature pairwise distances are calculated between corresponding Cα atoms in F_l. Next, each occurrence of a pair $F_i, 1$, $F_i, 2$ at a distance d is counted in G_i. This three dimensional box, is now a degenerate representation of the actual binding site, but stored in a rotationally independent manner. The acronym RIPr will be assigned to designate this **R**otationally **I**ndependent **P**ocket **r**epresentation.
In order to avoid issues due to discretisation of C_α to C_α distance into separate bins a filter of (1,2,1) was applied only in the distance dimension over each occurrence of a feature pair. This allows smoothing the discretised distance resulting in less sensitivity to small distance variations on C_α positions.
RIPrs are degenerate, as they *(i)* translate very precise geometric information and constraints of a binding site to pairwise distances of alpha carbons only and (ii) don't specify the exact count of a given feature, but reduce this information to a binary fingerprint. The second reduction of precision on the exact number of features hasn't been used so far by other methods to the best of our knowledge.

Advantages of a RIPr: The reduction of a binding site to a RIPr has limitations but also advantages, especially compared to other fingerprint methods. Compared to methods like FuzCav [Weill and Rognan, 2010], the size of a RIPr is adaptive to the number of features and the size of a pocket. Thus, instead of systematically using 4833 integer fingerprints, RIPrs can be of variable size (on average 1836±680). Given the number of cavities to analyse (hundreds of thousands), reducing memory usage and disk access is beneficial. A further advantage of RIPrs is explained in more detail in the paragraph "Averaging RIPRs".

Similarity between RIPrs: Finally, lets consider two pockets that have to be compared to each other. Both are transformed to a RIPr. These can then be used for direct comparison between each other. In order to do so a measure similar to a maximum mutual overlap between feature to feature to distance pairs is calculated. This overlap is then normalized by each RIPr

under investigation to derive two similarities. The average of both similarities is considered to be the fina similarity. This is tentatively explained here: Let F_i be the set of all fingerprints under consideration in space i (RIPr i), D the set of distances. Then we define

$$G_i : F_i \times F_i \times D \to \mathbb{N} \tag{3.7}$$

which maps each $a \in F_i, b \in F_i$ and $d \in D$ to the number of pairs with fingerprint a and b at distance d.

$$G_i(a, b, d) = \text{The number of pairs with a and b at distance d}$$

Now consider two spaces i and j, corresponding to RIPr_i and RIPr_j. For these let $F_{ij} = F_i \bigcap F_j$ be the set of mutual binary feature fingerprints. Then we define the following numbers

$$S_i(j) = \sum_{a \in F_{ij}} \sum_{b \in F_{ij}} \sum_{d \in D} G_i(a, b, d) \tag{3.8}$$

$$T_i = 4 \cdot \left(|F_i|^2 - |F_i| \right) \tag{3.9}$$

$S_i(j)$ is the number of pairs in RIPr_i also present in RIPr_j. Note that $S_j(i)$ is, in general, different from $S_i(j)$.
Then let $O_{ij} : F_{ij} \times F_{ij} \times D \to \mathbb{N}$ be the following map

$$O_{ij}(a, b, d) = \min\{G_i(a, b, d), G_j(a, b, d)\} \tag{3.10}$$

for $a \in F_{ij}, b \in F_{ij}$ and $d \in D$, which yields the number of mutual pairs in both RIPrs. Now as a measure of similarity between the two RIPrs:

$$W_i(j) = \frac{S_i(j)}{T_i} - \frac{1}{S_i(j)} \left(\sum_{a \in F_{ij}} \sum_{b \in F_{ij}} \sum_{d \in D} G_i(a, b, d) - O_{ij}(a, b, d) \right) \tag{3.11}$$

$$REF_i = \frac{1}{T_i} \left(\sum_{a \in F_i} \sum_{b \in F_i} \sum_{d \in D} G_i(a, b, d) \right) \tag{3.12}$$

$$SIM(i, j) = \frac{1}{2} \left(\frac{W_i(j)}{REF_i} + \frac{W_j(i)}{REF_j} \right) \tag{3.13}$$

Here the global similarity SIM(i,j) is an average value of two similarities of RIPr_i versus RIPr_j and vice-versa. However, if both pockets are of very distinct size and for instancee RIPr_i is a submatch of RIPr_j, then averaging both similarities can yield to a substantial drop in the final similarity. Thus both mutual similarities can also be used to detect global matches and sub-matches of pockets.

Averaging RIPrs: An interesting advantage of the RIPr is that its formulation can be easily handled. Actually, simple mathematical operations can be carried out on these objects. This turned out to be of particular use to reduce the pocket space to analyse. A major difficulty for creating the pocketome is combinatorial. Performing an all against all comparison of 350.000 cavities against each other is timely very costly and also needs large amounts of disk space. The RIPrs can be used to reduce this number of cavities by following the thought that, if two RIPrs are extracted from the same protein and are very similar (SIM(i,j) above a given threshold) then one can average them to an average RIPr representing two or more pockets in one single RIPr.

Validation

A dataset of pockets derived from PDB-Bind [Wang et al., 2004, 2005] and the pocket database presented in section 3.4 were used for a first validation attempt of this method. This PDB-Bind core-data-set contains 231 cavities. Several of these 231 cavities can be situated on the same protein, but on different structures. I established a non redundant list of proteins with their cognate binding site and considered it for further processing if the binding site was encompassing only a single protein chain.
Next, the binding site of these proteins was identified in the pocket database retrieving the corresponding pocket cluster. More precisely, using one protein and the cognate binding site as seed, all other occurrences of the same pocket on other structures are retrieved. If no pocket cluster was found due to the lack of structures or inconsistencies in clustering, the protein was discarded from the data-set. Furthermore, small pockets (less than 60 alpha spheres) were discarded. This procedure resulted in set of 59 proteins containing 1 cognate binding site each. As several crystal structures are known for each protein, a total of 2809 cavities corresponding to these cognate binding sites on different crystal structures are considered for this study.

Statistical aspects

In order to determine the ideal cut-off or threshold of the continuous similarity measure presented in equation 3.13 the F-measure was used. One can define one binding site as query (1) and all other binding sites as decoys (0). The predictive power of the similarity measure can be assessed via retrieval of all cognate binding sites from structures of the same protein with higher scores than the rest. This retrieval can be performed on various levels of a scoring threshold (upon which one would define a hit to be the same pocket as the query). The F-measure is the harmonic mean of precision and recall of a classification method.

$$Precision = \frac{TP}{TP + FP} \tag{3.14}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.15}$$

$$F = 2\frac{Precision \times Recall}{Precision + Recall} \tag{3.16}$$

Incrementing the similarity measure from its minimum to 1 by increments of 0.002, the maximum F-measure indicates the best score threshold. An F-measure of 1 indicates best classification performance while 0 indicates lowest performance.

3.5.4 Results

As the work undertaken here is new in the sense that pocket comparison is performed on automatically defined cavities on the whole protein surface, various validation studies have been performed to ensure that *(i)* the pocket comparison method allows to identify the same pocket among different structures and that *(ii)* the method allows to find similar pockets on other proteins with known relation.
The pocket similarities of all against all pockets are calculated for all cavities of the validation data-set derived from PDB-Bind. The resulting similarity matrix was further used to perform the following analysis.

Finding the similarity threshold

A central point in development of a continuous similarity measure is to know above (or below) which threshold a result is either positive or negative. Here, the pocket similarity measure varies from < 0 to 1, where 1 means a 100% identity between two cavities. Knowing that our data-set contains several times the same cavity but on different structures we can measure above which similarity threshold the method is able to retrieve most pockets of the same protein among all 2809 cavities in the data-set using one cavity as query. Figure 3.27 shows average F-measures obtained on the whole data-set (A) as well as two extreme examples (B,C) to highlight the variability one encounters in automatically predicted cavities. The F-measure is maximised (0.75 ± 0.15) at a score around 0.4 considering all cavities. The variability is calculated using each cavity in the data-set as query to retrieve all cavities in the known cluster of the same pocket on different structures.
An example for a pocket that is well identified by fpocket without major variations is the binding site of β-glucosidase A (figure 3.28 B). Thus the variation on similarity predictions is minor around the maximum F-measure (figure 3.27 B). On the contrary, β-lactamase is an example of a very open and solvent exposed cavity (3.28 A) yielding variable pocket predictions with fpocket (figure 3.27 C). As a result, average and also individual F-measures are penalized.
From this analysis two observations can be made: *(i)* the pocket similarity measure is capable of retrieving varying, but similar pockets detected on other structures from a pool of binding sites, *(ii)* the similarity threshold for retrieving similar cavities appears to be approximately 0.4. The exact value of the threshold can be discussed especially seen the layout of the analysis here. However, it gives a sound reference for later analyses. The threshold is very likely to be under-estimated. Given the fact that the layout of this experiment considers that binding sites of one protein like trypsin are fundamentally different to FXa while retrieving all trypsin binding sites can be a source of error. Indeed several serine proteases are part of the data-set and are similar

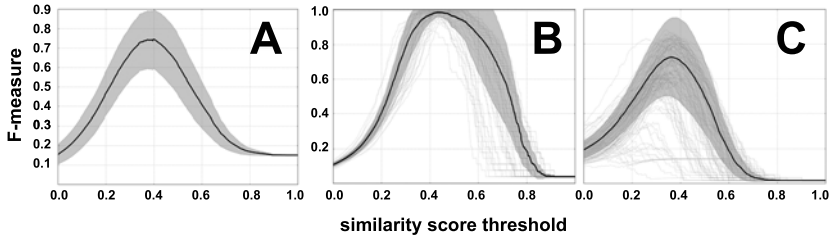


Figure 3.27: F-measure versus different similarity thresholds. Black thick curves: average values, grey shaded area: average ± standard deviation of values,**A:** Average F-measure for all cavities on all proteins in the data-set, **B:** Average F-measure for β-glucosidase A, dashed lines are results for individual cavities, **C:** Average F-measure for β-lactamase (y axis the same as in B), dashed lines are results for individual cavities.

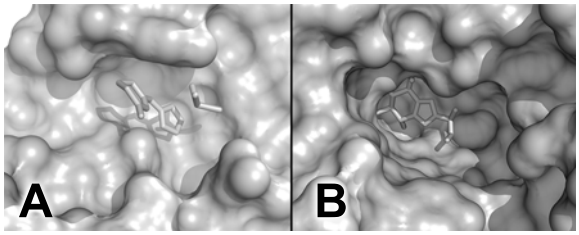


Figure 3.28: Example binding sites of **(A)** a solvent exposed binding site of β-lactamase and **(B)** a buried pocket in β-glucosidase

between each other. Thus they can introduce noise and lower the perceived optimal threshold. As discussed earlier, another source of error likely to reduce the perceived best threshold is the fact that we consider that all conformations of a cavity and all automatic pocket detection results should be treated as a consistent ensemble representing equally well the same pocket. Pocket conformations and especially detection results can vary substantially, reducing the perceived similarity and thus reducing the best perceived similarity threshold extracted from this analysis.

Validation all against all

An all against all comparison of every pocket in the data-set has been performed with the method described previously. Usually, results of such a study are represented as a dendrogram, but networks are becoming more and more common in the scientific literature to represent clustering results and we have chosen this representation too.
One node in such a network is a pocket, and two pockets (nodes) are connected by an edge if the similarity between them is above a given threshold (0.45 here). The advantage of such network representations is that powerful layout engines can be used to naturally dispatch the information in a 2D or 3D representation. Thus visual analysis is more straightforward, if the network is not too complex. Results extracted from the analysis carried out here are shown on the following

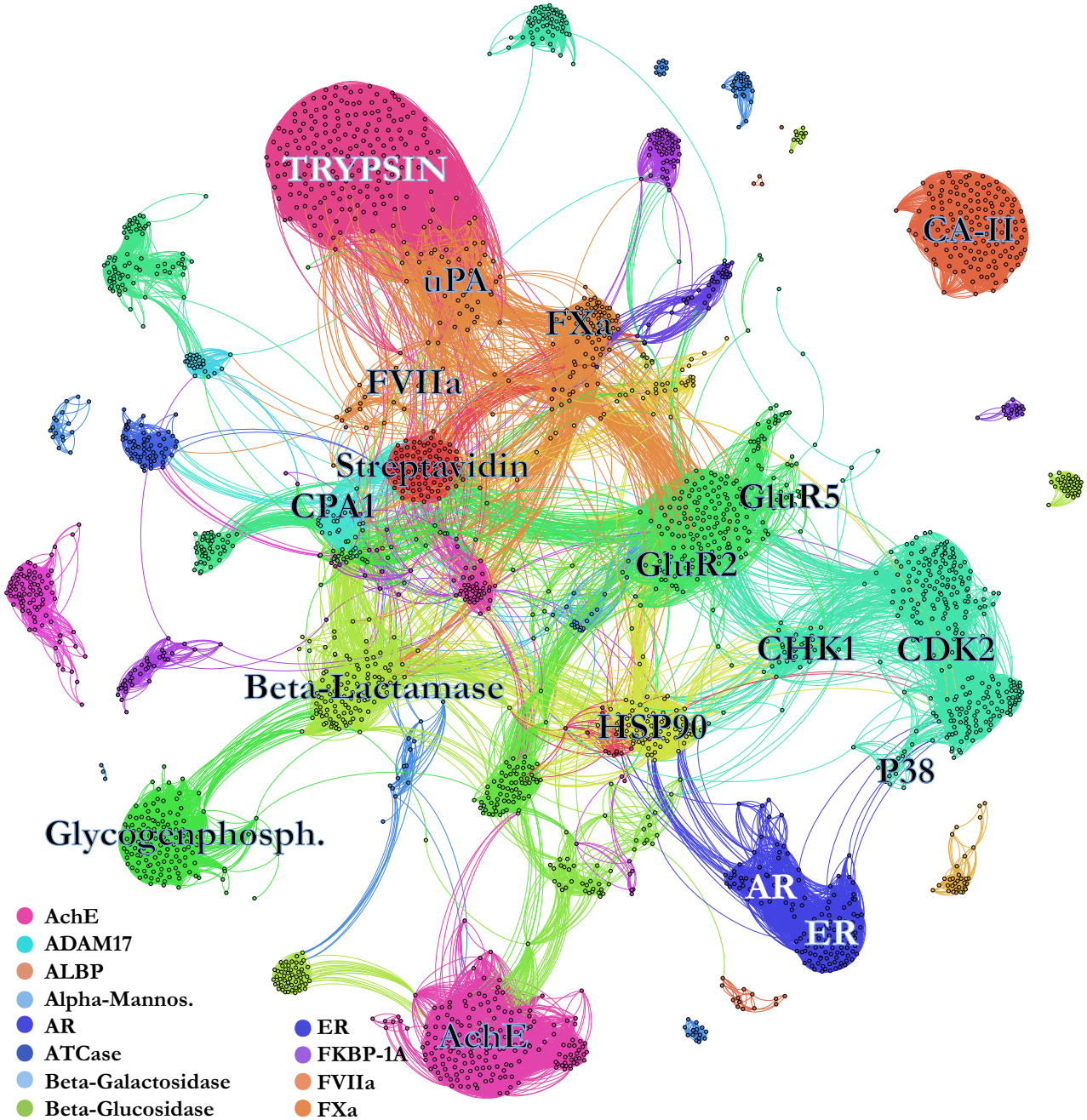
page. Nodes have been coloured by protein name and most important clusters of binding sites of proteins are labelled on the network.

Same pockets on different structures As a first important result it can be observed that all binding sites from the same protein usually cluster together into rather dense clusters. The density of such a cluster is directly related to *(i)* the number of connections one pocket has to all others of the same cluster of pockets, and to some extent *(ii)* the number of outbound connections to other pocket clusters from other proteins. This result is in good agreement with the previous analysis using the F-measure, showing good predictivity identifying the same pocket on different structures. While this is of limited interest for the pocketome project itself it has important technical and theoretical implications. First, it implicitly validates the pocket similarity measure and, second, it could be used to reduce the size of the pocketome efficiently.

Similar pockets on related proteins More challenging to analyse are results on interrelations of binding sites of similar as well as unrelated proteins in this graph. The data-set contains proteins that are related. For example one can find several kinases (CDK2, CHK1, P38) as well as endopeptidases and glutamate receptors. In this comparison a clear cluster for endopeptidases can be pointed out (trypsin, urokinase-type plasminogen activator - uPA, factor Xa and factor VIIa). A part from sharing a common function these proteins have similar sequences and folds. While previous validations proved that the pocket comparison protocol can identify the binding site on the same proteins but different structures, these results suggest that binding sites with minor variations are also found to be similar. This is furthermore highlighted by the cluster formed by kinases Chk1, CDK2 and P38. Interestingly, Chk1 is not part of the same kinase sub-family as CDK2 and P38, but similarities between all ATP binding sites are found. Another well known relationship exists between the androgen and estrogen receptor (AR, ER), both nuclear hormone receptors. The hormone binding sites used in this evaluation form a tight cluster showing the high structural similarity between the two pockets. The results shown so far can be of high interest when analysing proteins with related function or significant sequence similarity. Thus, the pocket similarity measure can be used to help and guide protein function prediction efforts in combination with other methods.

Similar pockets on unrelated proteins The analysis of the all against all comparison also reveals connections of proteins that are unrelated in function, sequence and fold. One of such a relation exists between HSP90 and CHK1. Although HSP90 is known to bind ATP in its active site, it has a very different topology compared to kinases like CHK1.

Validation all against all



- AchE
- ADAM17
- ALBP
- Alpha-Mannos.
- AR
- ATCase
- Beta-Galactosidase
- Beta-Glucosidase
- Beta-Lactamase
- Beta-Mannosidase
- Beta-secretase
- CA-II
- Cationic trypsin
- CDK2
- CHK1
- CPA1
- DHFR
- Endoglucanase
- Endothiapepsin
- EPSPS
- ER
- FKBP-1A
- FVIIa
- FXa
- GH84
- GluR-2
- GluR-5
- Glycogen phosphor.
- GSTP1-1
- HMGCR
- HMT
- HSP90
- Leucyl Aminopept.
- MME
- MMP-8
- Neuraminidase
- NMD-R1
- P38
- Panc. alpha amylase
- Pantothenate
- Penicillopepsin
- PKA C
- PLA24
- PNMTase
- PNP
- Polyprotein
- PTP1B
- Queuine tRNA r.
- RNase 1
- Scytalone dehydratase
- Streptavidin
- Stromelysin-1
- Thermolysin
- Thymidilate Synth.
- TIM
- Transthyretin
- uPA

Despite this fact, similarities are found between the two binding sites, known to bind ATP. More striking however are relations between three occurrences of the binding site of HSP90 and nuclear hormone receptors (NHR).
On first sight, known NHR binders share few common characteristics with HSP90 inhibitors or ATP. A structural superimposition of both binding sites shows that the pocket lumen is of approximately the same size and shape. Furthermore both cavities share some common hydrophobic patches. Interestingly, in a recently published work by Zhao et al., 4-hydrotamoxifen (4OHT), a known estrogen receptor binder, was systematically docked using computational docking software into 4403 known redundant binding sites of different proteins. Within the top 15 targets identified for 4OHT using this approach, 12 corresponded to estrogen hormone binding sites. Furthermore, 3 ATP binding sites of HSP90 [Zhao et al., 2010] are found. Given these unexpected results the authors verified the binding site similarity between the estrogen receptor hormone binding site and the HSP90 ATP binding site using SuMo [Jambon et al., 2003], a well known pocket comparison method. No significant similarity was found. In subsequent experimental analyses the authors show that 4OHT and tamoxifen activate HSP90 activity. Unfortunately, the exact interaction site of 4OHT on HSP90 could not be verified experimentally. This encouraging result shows how the introduction of uncertainty about atom positions can help to identify unforeseen relationships that cannot be spotted using more restrictive algorithms such as SuMo.
Despite these results several other links between unrelated proteins can be observed (streptavidin, glutamate receptor) where verification needs more in depth analysis and experimental proof. Analysing similarities on unrelated proteins is a challenging task especially if a putative relation is not known yet. However, discovery of relationships between unrelated proteins is of high interest in drug safety, but also in drug design itself. In the following paragraphs, several examples show the putative applications of the pocket comparison method presented here.

Screening the pocket database

Until now it has been shown that the method is able to identify similarities between pockets above a certain threshold. However, it is not known yet if higher similarities relate to higher likelihood to find the same pocket and lower similarities to find similar pockets on unrelated proteins. To investigate the behaviour of the score in this regard, a PDB wide cavity comparison was performed using one or multiple occurrences of a known binding site of a set proteins as query. Resulting similar cavities were ordered by score and for known binding sites among them (with cognate ligand) the cavities were classified in *(i)* the same cavity, *(ii)* a related cavity in a related protein, *(iii)* and unrelated cavity but known to bind ligands that share similarity with cognate ligands of the query pockets and *(iv)* unrelated cavities on supposedly unrelated proteins. The classification was done by hand and cavity by cavity, thus allowing only assessment of a limited amount of targets and resulting cavities given the important number of results.

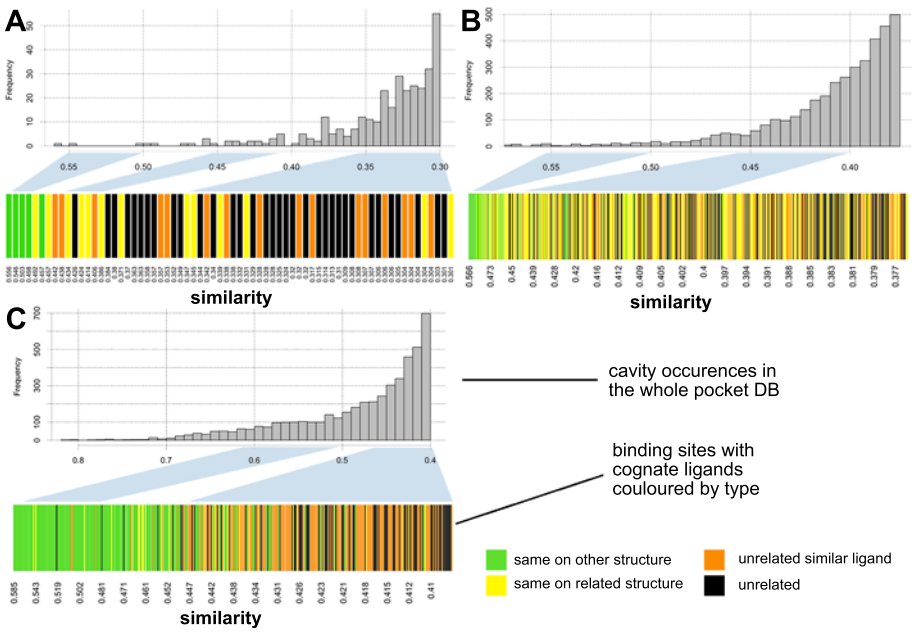


Figure 3.29: Binding site screening results for **(A)** Penicillopepsin, **(B)** Casein kinase 2 and **(C)** Acetylcholinesterase. Histrograms show all cavities (occupied or empty) found to be similar to query pockets. The coloured bar plot shows hits (one line corresponds to one pocket) of binding sites containing ligands. These hits have been classified by hand according to the relation to the query pocket.

Classifying the same pockets on the same protein but other structures as such is straightforward, thus there is no error to be expected from such a by-hand classification. Second, classifying a pocket as of being on a related target can yield a few errors. In the examples shown here, related targets were all kinases for example, if a kinase was used as query protein. For acetylcholinesterase, only the butyryl cholinesterase was considered to be related.
Last, the classification of an unrelated protein binding a similar ligand is highly subjective in this study. I would like to emphasize on the fact that this limitation is known and an automated and systematic way of classification has to be found. Here ligands were deemed similar if their global scaffold or shape were considered to be similar after ocular inspection. Despite these limitations the manual classification should prove sufficiently accurate to see global trends in the resulting screen.
Figure 3.29 illustrates results obtained during this screen. First, the penicillopepsin active site (one single structure) was compared to all cavities in the pocket database. Results shown on figure 3.29 A show that only few hits have been found. The pocket comparison method is however capable of retrieving the same binding site on other structures as highest scoring pockets (green bars on the left). While lowering similarity related binding sites are found (yellow bars), like endothiapepsin, penicillin binding protein 2X and saccharopepsin

binding sites. It is also notable in this example that below the similarity threshold of 0.4 a significant amount of pockets from apparently unrelated proteins are found.
Second, fourteen casein kinase ATP binding sites have been used as query against the pocket database. Figure 3.29 B shows the results of this screen. Here again the pocket comparison method is able to predict the same pockets (green) with higher similarities (most of them above 0.45). However, the fact of using more query cavities enables us to retrieve more results than in (A). The kinase space is also structurally very well represented in the PDB, explaining the abundance of identified similar cavities. Several related binding sites (yellow) can be found on other Ser/Thr kinases. Thus, the most similar cavities identified are the ATP binding sites of Pim-1, CDK2, c-Abl, adenylate kinase, rhodopsin kinase, rho-associated protein kinase, tyrosine protein kinase Fes/Fps, Braf, the insulin receptor and Chk1. Therefore, various kinase families covering Ser/Thr as well as Tyr kinases are found with similarities above 0.45. This result is in stark contrast with a study on functional classification of protein kinase binding sites using Cavbase [Kuhn et al., 2007]. In this work the authors use Cavbase to compare a set of 285 kinase binding sites and cluster them by similarity. It was found that this procedure was able to cluster generally protein kinases from the same family together. Results obtained here, screening only for Casein kinase 2 suggest that at least for this particular target the method is able to retrieve structurally similar kinases, but of a broader scope than Cavbase. This difference is very likely to be due to differences in the pocket representation (level of uncertainty) between the two methods.
Interestingly among the first 10% of hits towards casein kinase one can find unrelated proteins binding similar ligands, like a hypothetical dUTPase, serum albumin and a GFP like non-fluorescent. While dUTPase can be expected to be found as it is co-crystallized with nucleotides, the later two hits seem less obvious at first sight. Serum albumin is known to bind a large variety of molecules like DNA, fatty acids, hormones and also drugs [Sjöholm et al., 1979]. However, 2 tyrosine kinase inhibitors were found to bind to serum albumin [Chandrasekaran et al., 2010]. The similarity with GFP like and other green fluorescent proteins is a probably the less obvious. It is a very good example of the level of uncertainty of the method. Comparing the ligand found in the crystal structure of GFP (PDB code: 1oxf) one can observe that the ligand shares the overall topology of adenosine, while replacing adenin by a amino-indol ring and the ribose by an imidazole. This result seems rather counter intuitive but it reflects a wished behaviour of the pocket comparison algorithm.
Given the multitude of similar pockets found with respect to Casein kinase 2 they cannot be discussed here in detail. Shortly, several other kinases are found at lower similarities but also unrelated proteins binding ATP & ADP. Interestingly, also here we can identify HSP90 as being similar while not sharing the fold nor the function of Casein kinase 2.

In a last test case, 32 occurrences of the acetylcholinesterase binding site were screened and results are shown on figure 3.29 C. An important amount of acetyl-

cholinesterase active sites are gathered with high similarities (green bars on the left). Also with very high similarity, the active site of butyryl cholinesterase, a closely related protein to acetylcholinesterase, is found. A first unexpected hit is the renin binding site. However, tacrin an acetylcholinesterase inhibitor and derived compounds are very similar to a part of the carbazole of the ligand in PDB structure 2g1o of renin. This ligand is known to bind with an IC50 in the submicromolar range. Here again, the pocket comparison method identifies sub-matches of the binding site, as the whole renin binder would not fit into the acetylcholinesterase active site. More intriguing is the appearance of similar cavities binding nucleotide like ligands. Unfortunately in literature, binding of such ligands is not reported on acetylcholinesterase or these are reported not to bind. However it has been published that cycloSal pronucleotides bind and show activity on butyryl-cholinesterase [Meier et al., 2004] already shown to be similar to the query protein. This is a very interesting result, showing that the pocket comparison can *(i)* identify pockets that are further away in similarity, with lower similarities, and identify relations between pockets that do not seem closely related but are through an intermediate.
In conclusion these results indicate that screening for similar pockets among 350.000 cavities yield positive hits on the same protein, allow to retrieve similar pockets on known related proteins with similar functions and most interestingly pockets on proteins without apparent relation but with ligands sharing similar scaffolds or global topologies. This shows that the fuzziness of the method allows to relate pockets that are different from a structural point of view, but share global shape and distribution of physicochemical features inside the pocket. The analysis also pointed out a difficulty for interpreting, navigating and representing these results. Interpretation is especially hindered when known binding sites appear without apparent relation. Given the ultimate objective to relate all pockets with all pockets regardless of the fact if they are known to bind small molecules an easy to use navigation tool needs to be developed.

Creation of the pocketome

What is the similarity between all pockets found on all structures in the PDB? Although data and methodologies exist to answer this question, currently no freely accessible tool exists to answer it. While theoretically it can seem straightforward to tackle this problem, the complexity and amount of data renders this task practically very difficult. While previous attempts in the same direction focused on known binding sites with cognate ligands, no systematic pocket comparison between binding sites and uncharacterised pockets has been performed yet, although around 80% of all pockets in the PDB do not contain a small molecule. The main objective of this work is to get a first insight into interrelations between all cavities in the PDB, albeit known binding sites or uncharacterised pockets, eurcaryote or pro-caryote.
Practically this means to perform a pocket comparison of around 350.000 cavities with each other resulting in 61.0 10^9 comparisons. Currently one comparison can be done in the millisecond range situating the cavity comparison

algorithm among the fastest in the field. If only a single CPU machine was used calculation time can be estimated to be of approximately 2 years. Using distributed computing on 128 CPU's calculation time could be downscaled to a week, however this ecologically very questionable strategy was not followed here. Instead, prior to calculation, a step of reduction of data redundancy was performed. Information in the PDB is very redundant, so are pockets on protein structures. As already seen on previous results the pocket comparison method is able to identify the same pockets on different structures. Thus the method can be used to detect clusters of very similar pockets (i.e. similarity threshold of 0.55) on one protein. These clusters have been identified and were used to derive an average pocket representing all individual pockets that are part of the cluster. This averaging of pockets can be easily achieved using RIPrs as explained previously. After averaging all pockets into average pocket clusters (APC) only non singleton clusters are considered for entering the final comparison versus all other clusters. Finally, only 28.000 APCs needed to be compared with each other to establish a similarity matrix as basis of the pocketome.
Another important argument for using pocket averaging prior to an all against all comparison is the obvious reduction of the results. Practically, handling, storing and analysing 400 million pairwise similarities is more tractable than 150 fold more.
Ultimately all APCs have been compared between each other. The resulting similarity matrix was analysed and similar pairs of APCs written to flat files if the similarity between them was above 0.4.
The pocketome derived previously is highly complex and contains a multitude of interrelations. To help analysing this data here again a network or graph based solution has been considered. In such a solution one APC corresponds to one node in a graph. An edge is traced between two nodes if they have a similarity above 0.4. This graph is augmented with various information coming from the pocket database. For example, if ligands are co-crystallised in pockets composing an APC, a ligand is added as a node to the graph and connected via an edge to the APC. Different attributes are stored onto both ligand and APC nodes, like protein and ligand names, Smiles and Uniprot accessions, organisms etc. Such a network or graph can then be easily distributed and used within popular graph visualisation software. It should be noted that these results show for the first time a map of interrelations between pockets and ligands associated with pocket-pocket similarities including unknown and uncharacterised pockets for the whole PDB.

Navigating the pocketome

As stated previously analysing the pocketome is highly complex and out of the scope of the work presented here or the work one single person could do. To ease the analysis, the pocketome can be visualised using Cytoscape [Shannon et al., 2003], a reference software for complex network visualisation and analysis in the systems biology community. Cytoscape is free and can be downloaded from `http://www.cytoscape.org`. The pocketome will be distributed as a col-

lection of flat files, but also a ready to use Cytoscape session is provided that can directly be loaded with Cytoscape. Figure 3.30 shows an example session

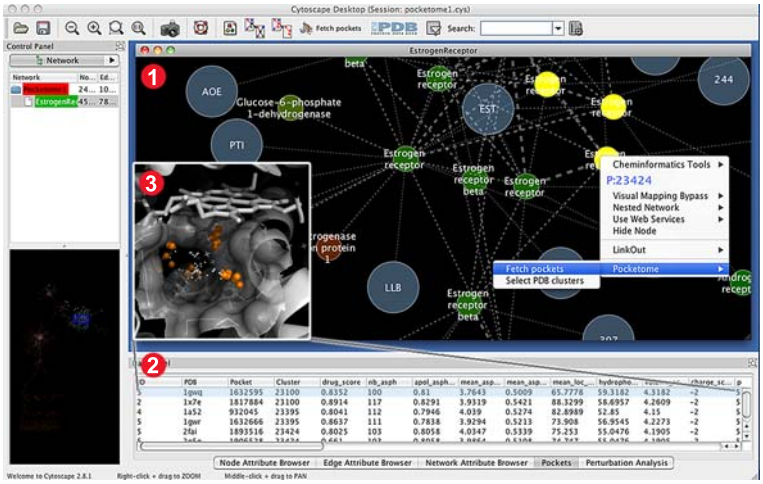


Figure 3.30: Graphical user interface of Cytoscape with the pocketome. **(1)** In the main window, nodes (average pockets and ligands) can be shown and coloured by various attributes. **(2)** By clicking on a pocket node, actual pocket details can be retrieved from the pocket database. A click on one particular pocket opens the structure and pocket definition in an independent PyMol session **(3)**.

using Cytoscape with the open pocketome. By default, no visualisation is activated as visualising the whole pocketome is very complex. The user has to focus on a particular protein or ligand. These can be easily selected using filters. Once done, all connected nodes (pockets and ligands) can be selected, thus extending the selection within the pocketome. Finally, the selection can be extracted as sub-network. Such a sub-network is easier to visualise and handle (part 1 of figure 3.30.

On figure 3.30 a part of the estrogen receptor sub-network is shown. All pocket nodes are interactive and upon a click the user can connect to the pocket database to retrieve which PDB structure and cavity corresponds to the APC seen in the graph. This is done via a purpose built plugin for Cytoscape, Cypockets, developed by Vincent le Guilloux. Details retrieved from the pocket database are loaded into the bottom of the window (part 2), showing PDB codes and pocket properties derived from fpocket. Each pocket line here is also interactive and again by a simple click on such a line, information from the pocket database is retrieved and the corresponding structure is opened in PyMol highlighting the pocket of interest.

The solution adopted here for distribution and visualisation of the pocketome allows to disseminate the results of the PDB wide pocket comparison very efficiently. The user does not need a local pocket database, or installed PDB. Furthermore, it has several advantages over popular web-server based publication of methods and results. Cytoscape and the pocketome can be used as standalone application and use is thus anonymous. The fact that Cytoscape

was chosen to represent the pocketome and navigate it allows the user to take advantage of other functionalities that can be directly linked to the pocketome. Such functionalities include extraction of known PPI networks data of a protein of interest. Furthermore the software is purpose built for complex network analysis and thus various functionalities available ease the navigation and analysis of the pocketome.

Applications

In this last section two application examples are shortly presented. These intend to show basically two intertwined aspects connected to the usage of the pocketome. In the first example it is shown how the pocketome can be used to characterise an uncharted binding site and identify putative ligands. In a second example an application very close to drug-repurposing or drug-recycling is outlined.

Characterisation of an unknown pocket in GSK3-β Glycogen-synthase-kinase 3 (GSK3) β, is a serine/threonine kinase that is part of the CMGC family together with well known kinases like P38α and CDK2. In an independent project we analysed all conserved pockets identified on the surface of this kinase. Among physico-chemical, druggability and conseration analysis of several pockets on this protein, a last analysis related to the pocketome was done.
A systematic screen of all pockets of GSK3-β was performed in a very similar way than the validation experiments performed on acetylcholinesterase, and casein kinase II presented earlier. While some of the several cavities investigated yielded no significant hits in the pocketome, one pocket showed intriguing results. Figure 3.31 B shows GSK3-β and the pocket investigated (grey circle). The screen for similar binding sites to this pocket yielded several results and other cavities in which known ligands bind. A striking result is its similarity to the ADP binding site on the Dha-kinase subunits DhaM and L. Interestingly, this binding site does not show the same architecture as known kinase ATP binding sites situated between a small beta strand lobe and a large alpha helix lobe.
Furthermore, this result is interesting, as already one known ATP binding site exists in this kinase. Here we identify that the investigated pocket resembles to an ATP binding site, a putative secondary ATP binding site. The superimposition between both, GSK3-β (green) and DhaM& DhaL shows several areas where backbone and side chain positions are highly similar between both pockets (circles). The arrow indicates a common termination of the helix lining the binding site. Naturally, seeing such important overlap in a sub-pocket arises the question if ATP or similar molecules could potentially bind there.
Unfortunately this question could not be answered via experimental nor computational techniques and is thus open for speculation. However, recently Shan and colleagues published an interesting work on very long MD simulations of self guided binding of known inhibitors to Src kinase. Results from this report are presented here in figure 3.31 A. Here red spots correspond to places where

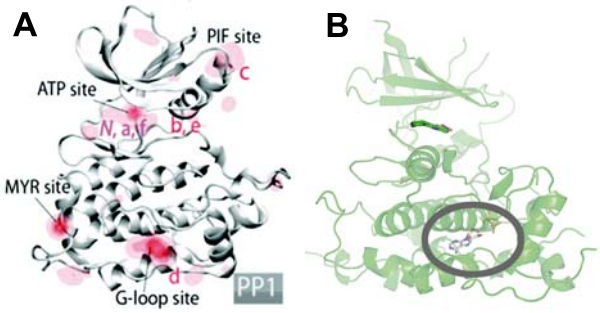


Figure 3.31: A: figure 1 taken from [Shan et al., 2011], showing hot spots identified via free MD of dasatinib with Src kinase. These results show a secondary binding site (G-loop site) for molecules known to bind to the ATP binding site. **B:** Structure of GSK3-β in the same orientation as Src kinase. The investigated pocket is contoured in grey.

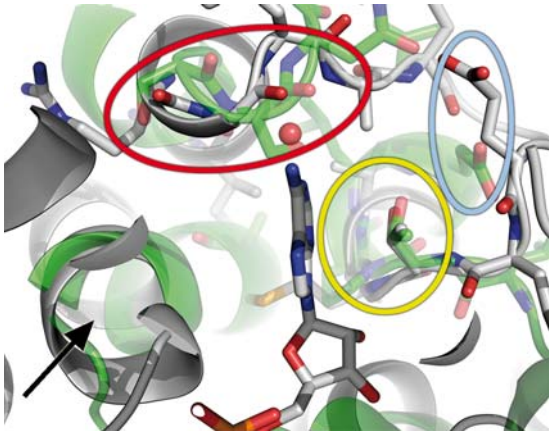


Figure 3.32: Superimposition of the investigated GSK3 pocket (green carbons) and the ADP binding site of Dha-kinase subunits DhaM and DhaL (white carbons). Intriguing matches are contoured by different colours.

desatinib was found to transiently stay on the pocket surface. It was found that next to the known binding site (ATP site) another location was identified (G-loop site). These findings are very much in line with what we can observe on GSK3-β. It can thus be postulated that the pocket that was investigated here might be able to transiently bind ATP or similar molecules.

These results arise important questions. How frequently can one find such a *pocket redundancy* on a protein and what role do these pockets have? The pocketome will be a resource of tremendous value to investigate such observations.

Discovery of cavities for drug-repurposing Once the pocketome navigation facilities within Cytoscape were finished I was able to take full advantage

of the easiness for exploring the pocketome. To validate the visualisation and pocket comparison I particularly focused on the sub-pocketome of nuclear hormone receptor binding sites.
Estrogen receptor (ER) was used as entry point into the pocketome and the network was increased by two levels to include similar pockets to ER pockets but also putative ligands of ER and these similar pockets. In the network, connected to the hormone binding site cluster of ER, Caspase 3 was found. After structural cross validation, the binding site identified is situated between the two peptides forming a dimer, p12 and p17. Figure 3.33 shows where the

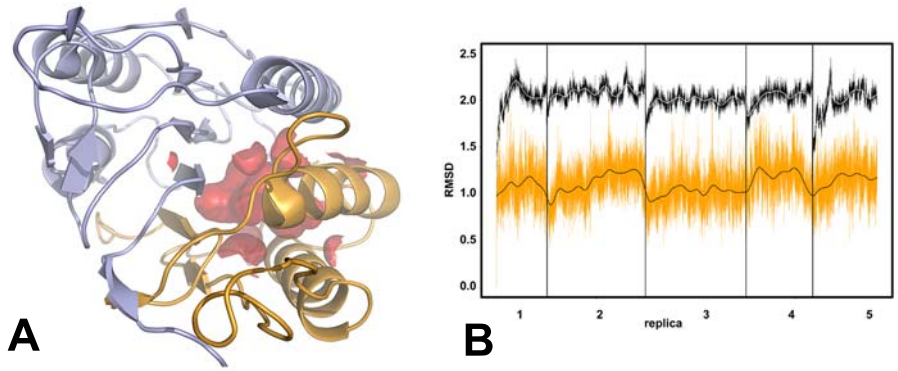


Figure 3.33: A: Overview of caspase-3. The p17 subunit is coloured in light blue and the p12 in orange. The active site is facing the observer. The newly identified cavity in caspase-3 is shown as red surface. **B:** RMSD of a compound identified via docking. 5 replicas of 20ns each were run to see if the protein (black curve) and the ligand (orange curve) are stable. RMSD calculated compared to the crystal structure with the docked ligand.

pocket is situated on the p12, p17 dimer. It is placed buried between both polypeptidic chains. Interestingly, this cavity was empty in the original crystal structure (1rhq). Next, it was investigated if the pocket is conserved among other structures of caspase 3. Systematically, small hydrophobic pockets are found in this same region on all other investigated structures, however the pocket identified in 1rhq is the biggest.
Based on the observation that this pocket is deemed similar to ER a very simple protocol has been adopted. We have taken estradiol as reference ligand for ER. Estradiol was used as input to the ZINC database [Irwin and Shoichet, 2005] to identify all molecules with a Tanimoto similarity above 0.7. The resulting molecules (around 500) were then systematically docked to this apo site using rDock [Morley and Afshar, 2004]. Importantly no relaxation of the binding site was previously done. Thus resulting binding poses were sterically very restricted. Despite these limitations, one compound was identified with a very promising binding pose.
This compound was then further assessed for stability in the pocket during molecular dynamics trajectories. A stable molecule in a flexible protein environment, surrounded by explicit solvent can give indications about the suit-

ability of a potential ligand for a target. The preliminary results from this MD analysis are shown on figure 3.33 B. Here the RMSD of the ligand (orange) and the protein (black) were tracked compared to the docking pose in the crystal structure. Five replicas of 20 ns each did not show any indication on possible instabilities of the molecule in this pocket.

These very encouraging results are currently experimentally validated with our collaborators. Given that the cavity seems naturally unoccupied, the fact of binding a small molecule and thus filling the cavity could have an important impact on the stability of the p12, p17 dimer. Caspase 3 has a very short half-life in its functional form (shown in figure 3.33 A). Thus stabilising the interaction between p12 and p17 could have a beneficial effect and alter the turnover of this enzyme that plays a central role in apoptosis [Porter and Jänicke, 1999].

Novelty

Here the first all against all comparison of putative binding sites on the whole PDB is presented. For the first time a pocket comparison is validated and carried out on automatically identified binding sites.
The method presented here uses a very abstract representation of the pocket allowing to compare structures that are very distant. The resulting pocketome is distributed in a novel, lightweight and intuitive to use manner and will be free to access for everyone.

Limitations

The pocket comparison measure shows very promising results in this priliminary study. However, validating such a method for relations that are non-obvious and where no experimental proof is available is very complex. Thus, a more systematic way to validate results from systematic pocket-comparisons has to be identified. While the fuzziness of the pocket representation allows retrieval of unforseen similar pocket pairs, it might hinder detailed studies of structural motives responsible for specificity of a given target towards a series of molecules.

A first unofficial version of the pocketome as Cytoscape session is available as supplementary material on the CD of the thesis. The pocketome folder on this CD contains a manual and the CyPockets plugin, such as that the members of the thesis tribunal can install, test and critically evaluate the resource. Note, that you need a working internet connection to run the CyPockets plugin.

Bibliography

R Albert, H Jeong, and AL Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–82, July 2000.

Andreas Bender, Daniel W Young, Jeremy L Jenkins, Martin Serrano, Dmitri Mikhailov, Paul A Clemons, and John W Davies. Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Combinatorial chemistry & high throughput screening*, 10(8):719–31, September 2007.

Salvatore Bongarzone, Hoang Ngoc Ai Tran, Andrea Cavalli, Marinella Roberti, Paolo Carloni, Giuseppe Legname, and Maria Laura Bolognesi. Parallel synthesis, evaluation, and preliminary structure-activity relationship of 2,5-diamino-1,4-benzoquinones as a novel class of bivalent anti-prion compound. *Journal of medicinal chemistry*, 53(22):8197–201, November 2010.

Imen Boukhris, Zied Elouedi, Thomas Fober, Marco Mernberger, and Eyke Hüllermeier. Similarity Analysis of Protein Binding Sites: A Generalization of the Maximum Common Subgraph Measure Based on Quasi-Clique Detection. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 1245–1250. IEEE, 2009. ISBN 978-1-4244-4735-0.

Appavu Chandrasekaran, Li Shen, Susan Lockhead, Aram Oganesian, Jianyao Wang, and JoAnn Scatina. Reversible covalent binding of neratinib to human serum albumin in vitro. *Drug metabolism letters*, 4(4):220–7, December 2010.

Howard J Feldman and Paul Labute. Pocket similarity: are alpha carbons enough? *Journal of chemical information and modeling*, 50(8):1466–75, August 2010.

Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–90, November 2008.

John J Irwin and Brian K Shoichet. ZINC–a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–82, 2005.

Martin Jambon, Anne Imberty, Gilbert Deléage, and Christophe Geourjon. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins*, 52(2):137–45, August 2003.

Esther Kellenberger, Claire Schalon, and Didier Rognan. How to Measure the Similarity Between Protein Ligand-Binding Sites? *Current Computer Aided Drug Design*, 4(3):209–220, 2008.

Sarah L Kinnings, Li Xie, Kingston H Fung, Richard M Jackson, Lei Xie, and Philip E Bourne. The Mycobacterium tuberculosis drugome and its polypharmacological implications. *PLoS computational biology*, 6(11):e1000976, January 2010.

Hiroaki Kitano. Towards a theory of biological robustness. *Molecular systems biology*, 3:137, January 2007.

Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature reviews. Drug discovery*, 3(8):711–5, August 2004.

Daniel Kuhn, Nils Weskamp, Eyke Hüllermeier, and Gerhard Klebe. Functional classification of protein kinase binding sites using Cavbase. *Chemmedchem*, 2(10): 1432–1447, 2007.

Sergei Maslov and I Ispolatov. Propagation of large concentration changes in reversible protein-binding networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13655–60, August 2007.

Chris Meier, Christian Ducho, Ulf Görbig, Robert Esnouf, and Jan Balzarini. Interaction of cycloSal-pronucleotides with cholinesterases from different origins. A structure-activity relationship. *Journal of medicinal chemistry*, 47(11):2839–52, May 2004.

S David Morley and Mohammad Afshar. Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock. *Journal of computer-aided molecular design*, 18(3):189–208, March 2004.

A G Porter and R U Jänicke. Emerging roles of caspase-3 in apoptosis. *Cell death and differentiation*, 6(2):99–104, February 1999.

Felix Reisen, Martin Weisel, Jan M Kriegl, and Gisbert Schneider. Self-organizing fuzzy graphs for structure-based comparison of protein pockets. *Journal of proteome research*, 9(12):6498–510, December 2010.

Yibing Shan, Eric T Kim, Michael P Eastwood, Ron O Dror, Markus A Seeliger, and David E Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–3, June 2011.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504, 2003.

I Sjöholm, B Ekman, A Kober, I Ljungstedt-På hlman, B Seiving, and T Sjödin. Binding of drugs to human serum albumin:XI. The specificity of three binding sites as studied with albumin immobilized in microparticles. *Molecular pharmacology*, 16(3):767–77, November 1979.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–80, June 2004.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The PDBbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–9, June 2005.

Nathanaël Weill and Didier Rognan. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *Journal of chemical information and modeling*, 50(1):123–35, January 2010.

Lei Xie, Li Xie, and Philip E Bourne. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics (Oxford, England)*, 25 (12):i305–12, June 2009.

Rongmin Zhao, Elisa Leung, Stefan Grüner, Matthieu Schapira, and Walid A Houry. Tamoxifen enhances the Hsp90 molecular chaperone ATPase activity. *PloS one*, 5 (4):e9934, January 2010.

Chapter 4

Conclusions

The work presented in this document has a manifold of possible applications. Before resuming conclusions of this work a short analysis of the impact this research had so far in the scientific community is presented.

Impact analysis

On the web

The fpocket project and all subsequent developments are presented via a website accessible on the internet. Such a website allows tracking traffic and the origins of this traffic. Figure 4.1 displays statistics tracked using Google Analytics and Sourceforge web-traffic tracking tools. Website access statistics are extracted for the URL http://fpocket.sourceforge.net. On this figure a clear

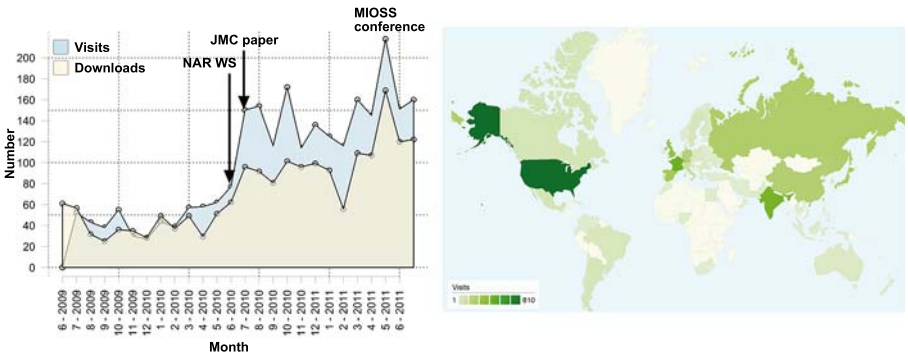


Figure 4.1: Number of visits on the fpocket website and fpocket downloads since june 2009. On the right, geolocalisation of visitors on the fpocket website. Statistics gathered with Google Analytics and Sourceforge.

trend of augmenting monthly web-traffic can be seen since the creation of the fpocket project. Furthermore, the publication of two fpocket related papers is indicated on the time-line and clearly shows the impact the papers had on

monthly visits and downloads of the program. On the right the geo-localisation shows that fpocket and connected projects yielded interest all over the world. Thus, it is not only a Spanish-French localised project, but universities and pharmaceutical companies mainly in countries like the US, Japan, Germany, United Kingdom, China, India and Russia seem to show interest in the project. Universities and institutions that have shown most interest in fpocket are the Moscow state university, the university of Barcelona and Paris Diderot, the university of Cambridge, the university of California Irvine, the Eberhard Karls Universität and several more. Among top visiting institutions are also pharmaceutical companies like Merck, Novartis and Hoffmann LaRoche Inc.
Analysing the number of calculations run on the fpocket online version hosted by the RPBS in Paris we noticed that in year 2010 alone around 4100 fpocket jobs, 500 mdpocket and 100 hpocket jobs were run. Interestingly, most job-requests came from China. Till today, fpocket has been downloaded nearly 2000 times.
Another interesting impact had the presentation of the fpocket project at the Wellcome Trust sponsored MIOSS workshop (Molecular Informatics Open Source Software) in Hinxton (UK). The fpocket project was invited as the only pocket prediction and characterisation tool to present itself together with well known and established projects in chemo-informatics, like RDkit, CDK, Openbabel, KNIME, Taverna etc. This was a first confirmation that the project is among the main-players in open-source pocket prediction.

In literature

Next to web-traffic, several papers were published related to fpocket, all of which are part of this thesis. The original fpocket paper is often cited (30 citations in Google scholar) and stated as highly accessed in BMC Bioinformatics. The method is also shown to still perform well in recently published benchmarks.
The article on druggability, that appeared in J Med Chem in summer 2010 had probably the highest impact in literature. The paper was in the Top 5 most accessed papers of the journal in Q3 of 2010. The importance of the contribution to the field made by that paper has also recently been acknowledged in a review published by a group at Pfizer around Enoch Huang (`http://www.sciencedirect.com/science/article/pii/S1367593111000895`). Here, this work was cited as being of outstanding interest to the field and the importance of collaborative data curation was highlighted, as well as the importance of polar atoms in druggable cavities.

Druggability predictions

- A novel data-set for training and validating structure based druggability methods has been derived
- A collaborative web-based platform called Druggable Cavity Directory was proposed to further develop and discuss the data-set with the com-

munity

- For the first time properties of polar atoms in binding sites are related to druggability. Polar atoms were found to make themselves available in druggable binding sites, although they are generally in an apolar environment
- A novel druggability score has been trained on the newly established dataset. This score was validated and compared to commercial state-of-the art prediction software
- The resulting score was implemented into the fpocket project and is distributed free of charge

Structure kinetics relations

- A special type of polar atoms was discovered to potentially play a central role in protein-ligand binding kinetics. These almost buried polar atoms (ABPAs) were found to act as kinetic traps on the protein surface
- A simple rule was derived saying that if the surface area of the polar atom is in the range of 2 to 10 $\mathring{A}$ and ΔA is negative then the atom is likely to behave as kinetic trap in apolar environments. This rule can be seen as the first general structure kinetic relationship.
- The concept was then validated on a concrete example of ligand optimisation. Two pairs of known HSP90 binders were analysed for their interaction with an ABPA in the binding site. It was shown that the interaction with the ABPA improves residence time of the ligand. Furthermore shielding an interaction with an ABPA is shown to be beneficial for the stability of the compound.

Protein pocket prediction on conformational ensembles

- A new protocol was developed to identify and characterise cavities on conformational ensembles of proteins
- The program called MDpocket is shown to produce visually easy to interpret pocket frequency and density maps
- Application use-cases are shown for transient channel predictions on Myoglobin MD trajectories. MDpocket is able to detect stable and transient channel openings
- The method is also shown to be of usefulness in the selection of protein conformations for ensemble docking. Here the mean local hydrophobic density of a cavity was shown to be predictive of a putative outcome of ligand docking on a given receptor conformation

- MDpocket is included in the fpocket project and is thus freely available

The pocket database & the pocketome

- A PDB wide cavity database was constructed. This database contains all pockets identified with fpocket, as well as their properties
- Supplementary data-sources were integrated to complete the pocket database. These sources include Uniprot, Kegg, PISA and PDB to Uniprot mappers.
- A protocol was derived to retrieve automatically druggable cavities on protein-protein complexes. These pockets are now further investigated for their ability to bind protein-protein interaction stabilisers.
- A comprehensive GUI was developed. This interface allows easy and intuitive navigation in the pocket database.
- A novel pocket comparison method was developed. This method represents the cavity as an abstract ensemble of interaction feature pairs.
- The pocket comparison method was validated for the first time on automatically detected pockets. It is shown to be very efficient and can be used to screen the whole pocket database, representative of the whole PDB.
- The method is shown to significantly enrich comparison results with similar binding sites.
- An all against all pocket comparison was performed on the whole pocket database. The result of this comparison was used to the derive the pocketome, relating similar pockets to each other and to known ligands
- Last, two promising examples of possible applications of the pocketome are shown
- The pocketome will be published free of charge and can be visualised in state-of-the art systems biology software Cytoscape.

Chapter 5

Appendix

Fpocket: An open source platform for ligand pocket detection

Vincent Le Guilloux,Peter Schmidtke*, Pierre Tufféry
**equal contribution, BMC Bioinformatics, 2009, 10(168)*

BMC Bioinformatics

BioMed Central

Software

Open Access

Fpocket: An open source platform for ligand pocket detection

Vincent Le Guilloux[†1], Peter Schmidtke[†2] and Pierre Tuffery[*3,4]

Address: [1]ICOA – Institut de chimie organique et analytique – UMR CNRS 6005, Div. of chemoinformatics and molecular modeling, University of Orléans, Orléans, France, [2]Dpto Fisicoquimica, Fac Farmacia, Univ Barcelona, Barcelona, Spain, [3]Molécules Therapeutiques in silico, INSERM, UMR-S 973, University Paris Diderot – Paris 7, Paris, France and [4]Ressource Parisienne en Bioinformatique Structurale, University Paris-Diderot, Paris, France

Email: Vincent Le Guilloux - vincent.le-guilloux@univ-orleans.fr; Peter Schmidtke - pschmidtke@mmb.pcb.ub.es; Pierre Tuffery* - pierre.tuffery@univ-paris-diderot.fr

* Corresponding author †Equal contributors

Published: 2 June 2009

BMC Bioinformatics 2009, **10**:168 doi:10.1186/1471-2105-10-168

Received: 23 March 2009
Accepted: 2 June 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/168

Abstract

Background: Virtual screening methods start to be well established as effective approaches to identify hits, candidates and leads for drug discovery research. Among those, structure based virtual screening (SBVS) approaches aim at docking collections of small compounds in the target structure to identify potent compounds. For SBVS, the identification of candidate pockets in protein structures is a key feature, and the recent years have seen increasing interest in developing methods for pocket and cavity detection on protein surfaces.

Results: Fpocket is an open source pocket detection package based on Voronoi tessellation and alpha spheres built on top of the publicly available package Qhull. The modular source code is organised around a central library of functions, a basis for three main programs: (i) Fpocket, to perform pocket identification, (ii) Tpocket, to organise pocket detection benchmarking on a set of known protein-ligand complexes, and (iii) Dpocket, to collect pocket descriptor values on a set of proteins. Fpocket is written in the C programming language, which makes it a platform well suited for the scientific community willing to develop new scoring functions and extract various pocket descriptors on a large scale level. Fpocket 1.0, relying on a simple scoring function, is able to detect 94% and 92% of the pockets within the best three ranked pockets from the holo and apo proteins respectively, outperforming the standards of the field, while being faster.

Conclusion: Fpocket provides a rapid, open source and stable basis for further developments related to protein pocket detection, efficient pocket descriptor extraction, or drugablity prediction purposes. Fpocket is freely available under the GNU GPL license at http://fpocket.sourceforge.net.

Background

In the recent years, in silico structure based ligand design (SBLD) has become a major approach for the exploration of protein function and drug discovery. It has been proven to be efficient in the identification of molecular probes, in investigation of molecular recognition, or in the identification of candidate therapeutic compounds (see for instance [1,2]). Whereas SBLD encompasses a wide range of aspects, one approach of importance is structure based virtual screening (SBVS). In SBVS, one searches, given the

structure of a protein, to dock candidate compounds to identify those likely to bind into a candidate ligand binding site (see for instance [3] and references included).

The identification and characterization of pockets and cavities of a protein structure is a key issue of such process that has been the subject of an increasing number of studies in the last decade. Several difficult aspects have to be considered among which: (i) the candidate pocket identification itself [4-26]. Here, one needs methods to identify and delimit cavities at the protein surface that are likely to bind small compounds. (ii) pocket ranking according to their likeliness to accept a small drug-like compound as ligand, for instance. Since often several pockets are detected at a protein surface, it is necessary to have some characterization of them in order to select the relevant ones. Although the largest pocket tends to frequently correspond to the observed ligand binding site (e.g. [18]), this rule cannot be generalised. Different studies have tackled this problem, see for instance [18,19,21,27,28]. It has in particular been shown that the use of evolutionnary information such as residue conservation helps re-ranking the pockets [19,21]. (iii) Last, but not least, there is often an adaptation – the so called induced fit – of the pocket geometry to the formation of a complex with the ligand (see for instance [29-32]). This last point creates several issues in terms of pocket detection – the pocket could or could not be properly detected in absence of ligand – and in terms of scoring since scoring functions are strongly dependent on the quality of the pocket identification and delimitation, but also are sensitive to conformational changes. Here, we focus on the primary but central aspect of candidate pocket identification from structure.

It is not easy to summarise the diversity of approaches that have been proposed so far for candidate pocket identification. Roughly, some are based on pure geometric analysis of the surface of the protein [4-15,18,20,22-26], whereas some others involve energy calculations [16,17]. Another way of distinguishing between the various approaches is to consider the detection algorithms. These can be classified as grid-based, and grid-free approaches. Grid based approaches [4,11,15-17,19,20,23] cover the proteins with a 3D grid and then search for grid points that are not situated within the protein and that satisfy some condition. For instance, POCKET [4], and the derived LigSite [11] search for protein-solvent-protein (PSP) events on the grids to identify pockets as positions enclosed on both sides by the protein. Pocket-Picker [23] uses a buriedness index to identify clusters of grid points likely to correspond to ligand binding pockets. Laurie and Jackson [17] position a methyl probe at grid points and calculate an interaction energy with the protein. An et al. [16] calculate a grid potential map of the Van der Waals force field using a carbon atom probe. Grid free approaches encompass (non exhaustive) probe (or sphere) based approaches as well as methods using the concepts of Voronoi diagrams. Sphere or probe approaches are based on the positioning of probe spheres at protein surface and to identify clusters of spheres having some property representative of candidate pockets. SURFNET [8,21] positions gap spheres between any pairs of atoms, reduces their radii so that they do not intersect any atom, and retains spheres with a radius more than a given threshold. PASS coats the protein using small probes positioned from unique triplets of atoms, and then identifies candidate pockets using a "burial count" – a number of protein atoms within a distance of the probe – to exclude convex parts of the surface. Iterative coating of remaining buried parts further allows the detection of "active site points" that represent the centres of potential pockets. More recently, both Nayal et al. and Kawabata & Go have proposed approaches using two different probe sizes to identify cavities. Small probes are used to identify a collection of positions at protein surface whereas large probes are used as a means to select the small probes located in depressions at protein surface. Among approaches related to Voronoi diagrams, CAST [13] and APROPOS [10], extract from the Delaunay triangulation of the convex hull the so called alpha-shape – a subset of the triangulation from which Voronoi vertices and edges outside the molecule are omitted. The commercial package SiteFinder [33] uses the concept of alpha spheres – spheres that contact four atoms and do not contain any atom (see concepts) – to identify cavities. Finally, Kim et al. [26] have recently proposed another approach based on the identification of "pocket primitives" from Voronoi diagrams.

In terms of availability, several of these approaches can be accessed via web servers (e.g. [34-36]), but very few packages are available for distribution. Some have been released as binaries (e.g. [14]), and for instance only the recently released PocketPicker [23] and LigSite(csc) [19] are available as open source softwares. There are a lot of research topics for which the availability of a free method can be of interest. Concerning this precise field, a part which is of major interest is development of scoring functions. These functions enable ranking of cavities when compared to each other. They are trained usually on descriptors of the binding pocket. Next, one has to assess rapidly the performance of these scoring functions. Still today, extraction of relevant pocket descriptors as well as assessment of scoring functions is an issue. One generally has to develop automatisation protocols for assessment. Available free tools performing these tasks might fasten discovery in computational binding site and drugability prediction. Besides, there are several scopes in which flexible software adaptation from source code might be required. For instance, the search for catalytic site pockets might differ from the search for protein-protein interaction effectors or carbohydrate-protein binding sites. Finally, speed remains an issue, in a context where the

pocketome size keeps increasing. In a general manner, the user should also be able to freely complexify the algorithm, in order to improve its performance and repropose the modifications freely to the scientific community. Thus, fast, accurate and high performing development based on a community willing to share their improvements might lead to a leading edge software package for pocket identification. PocketPicker makes one step in this direction. However, it was developed in Python and is specially adapted for visual purposes within PyMol. Thus PocketPicker seems adapted for punctual visual pocket detection, but not really adapted for large scale evaluations, especially due to execution speed limitations.

In this paper, we introduce a free pocket detection software called fpocket. It is based on the alpha sphere theory, an approach that relies on Voronoi tessellation that is among others the basis of the commercial software SiteFinder available within MOE from Chemical Computing Group [33]. It has several inherent advantages such as computational efficiency, the direct identification of the atoms of the proteins involved in a pocket, and promising possibilities to combine pocket detection and docking using a unified framework [37]. Using this approach, we propose a modular package to organise large scale pocket detection, descriptor extraction and benchmarking.

Implementation

Concepts

Fpocket relies on the concept of alpha spheres, introduced by Liang and Edelsbrunner [13] and also used by Chemical Computing Group in the SiteFinder software [33].

Briefly, an alpha sphere is a sphere that contacts four atoms on its boundary and contains no internal atom. By definition the four atoms are at an equal distance (sphere radius) to the alpha sphere centre. Alpha sphere radii reflect the local curvature defined by the four atoms: 4 atoms in a plane would correspond to an alpha sphere of infinite radius, and conversely, 4 atoms packed at the apex of a tetrahedron would lead to a value of radius close to that of the Van der Waals radius. For a protein, very small spheres are located within the protein, large spheres at the exterior, and clefts and cavities correspond to spheres of intermediate radii. Thus, it is possible to filter the ensemble of alpha spheres defined from the atoms of a protein according to some minimal and maximal radii values in order to address pocket detection. In practice, alpha sphere identification can be related to Voronoi decomposition of space: the centre of alpha spheres correspond to Voronoi vertices – points at which Voronoi regions intersect.

Once having identified a filtered ensemble of alpha spheres, another property of interest is that candidate regions of interest such as clefts at protein surface have larger occurrence of alpha spheres. Thus, the search for ligand pockets can be turned as the search for clusters of alpha spheres of proper radius. Finally, the knowledge of the spheres also comes with the identification of the atoms of the protein involved. It is thus easy to type the spheres according to some properties depending on the atomic types – such as for instance hydrophobicity – in order to filter the clusters. Conversely, knowing a pocket, it is also possible to extract properties for the atoms defining it.

Algorithm

The fpocket core can be resumed by three major steps. During the first step the whole ensemble of alpha spheres is determined from the protein structure. Fpocket returns a pre-filtered collection of spheres. The second step consists in identifying clusters of spheres close together, to identify pockets, and to remove clusters of poor interest. The final step calculates properties from the atoms of the pocket, in order to score each pocket.

Voronoi tessellation and alpha sphere detection

Voronoi tessellation is performed using the qhull package and more precisely the program qvoronoi [38]. Qhull's source code is freely available on http://www.qhull.org. Fpocket submits the heavy atom set for Voronoi tessellation to Qhull. In return Fpocket receives a set of coordinates of Voronoi vertices, atomic neighbours and vertex neighbours. This list of Voronoi vertices is then pruned according to two parameters: a maximum size of alpha spheres and a minimum size. Pruning the alpha spheres set by this maximum size and minimum size enables the elimination of solvent inaccessible alpha spheres and too exposed alpha spheres. Finally, only alpha spheres defined by zones of tight atom packing are retained and all the other alpha spheres are discarded.

Alpha spheres are then labelled according to the atom type they contact. Fpocket defines alpha spheres as apolar when they are contacting at least 3 atoms with a low electronegativity (< 2.8), typically carbons and sulfur in proteins. Subsequently, polar alpha spheres contact 2 or more polar atoms (typically oxygen or nitrogen).

Clustering of alpha spheres

This step has to be performed on several tenth of thousands of alpha spheres. Three different clustering steps are applied to the set of alpha spheres. The first one is a rough segmentation pass. In order to perform this step in a reasonable calculation time, fpocket uses the neighbour lists output from Qhull that indicates Voronoi vertices connected to each other by an edge. Fpocket checks if these interconnected vertices are close to each other and identifies a first set of clusters using a simple distance criterion. After this first pass, all clusters having only one sphere – generally large spheres situated at the protein surface – are

removed, and the centre of mass of each cluster is calculated. The next clustering step consists in the aggregation of clusters having proximate centres of mass. This way, small clusters of alpha spheres, especially on the surface are aggregated into one single cluster. Reducing complexity of an alpha sphere cluster on one single barycentre provides a rapid approach in order to group small clusters together, without performing a loop on all alpha spheres. Finally, a step based on a multiple linkage clustering approach is carried out in order to perform final fine clustering. During this step, all vertices of one cluster are compared to vertices of another cluster. If a certain number of alpha spheres of one cluster are near a certain number of alpha spheres of another cluster, both clusters are merged together.

After these three clustering steps, a pruning of uninteresting alpha sphere clusters can be performed. At this stage, small and essentially polar clusters can be dropped from the protein surface. User defined minimum number of alpha spheres and apolar spheres are used in order to influence removal of rather hydrophilic or small putative binding pockets. Note that this facility proposed to users is not used in the present study.

Characterization and ranking of the pocket

Last, clustered pockets were characterised in order to rank pockets according to their ability to bind small molecules. Note that the current ranking of pockets does not reflect drugability. It simply reflects the putative capacity of the pocket to bind a small molecule, that might be drug-like, but might also be a sugar, cofactor or coactivator. This rather basic but successful scoring scheme was derived using Partial Least Squares (PLS) fitting to some of the currently implemented pocket descriptors in fpocket.

Core programs

The fpocket package is made of three components: fpocket (Finding pockets) to perform the pocket identification, as described previously. Tpocket (Testing pockets) is provided in order to organise the benchmarking of the pocket detection algorithm over collections of structures, and dpocket (Describing pockets), designed to extract descriptors from collection of pockets from multiple structures. A flowchart of each is reported figure 1. Note, that the core of tpocket and dpocket is fpocket, exactly the same as the standalone fpocket program. Simply a layer of large scale statistical analysis was added to these two programs, in order to facilitate high throughput pocket detection and assessment of scoring performance.

Fpocket

Figure 1a illustrates the workflow of Fpocket (finding pockets), as well as the structure of the input and the output. This program will take as input a protein structure (PDB format) or a list of pdb files and return information about candidate pockets, numbered by rank. Fpocket will usually discard all atoms of the input file tagged as hetero atoms (including solvent and ligands). Nevertheless, cofactors like hemes should be kept during cavity detection, as they are usually part of the functional unit of a protein. Thus, fpocket maintains a list of cofactors

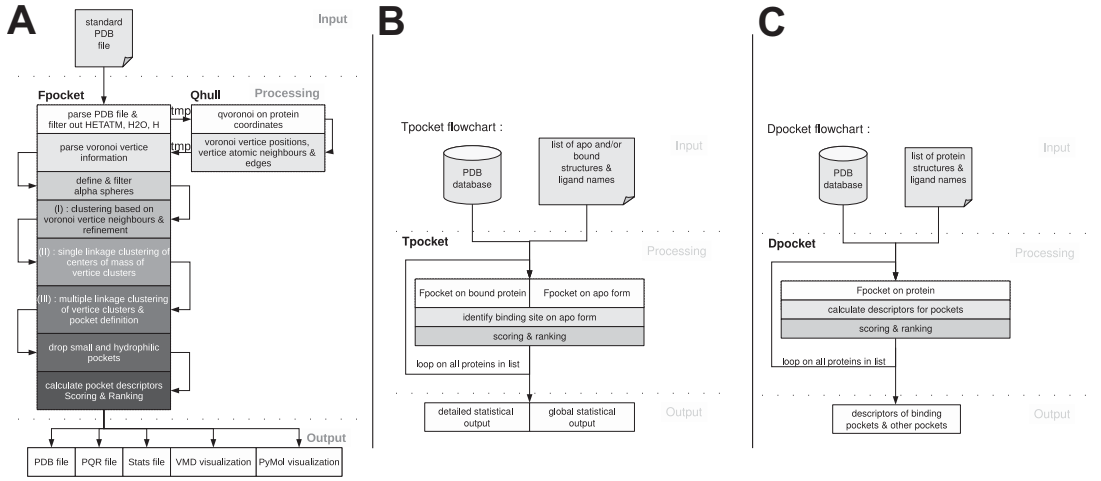


Figure 1
Fpocket (A), Tpocket (B) and Dpocket (C) flowcharts.

accepted as part of the protein during pocket detection. The algorithm is controlled by several parameters that can be adjusted by the user:

- Alpha sphere filtering parameters are related to the minimum (resp. maximum) size of alpha spheres: the minimum (resp. maximum) size an alpha sphere might have during alpha sphere docking on Voronoi vertices. Alpha spheres beneath (resp. above) this size are discarded from clustering.

- Alpha sphere clustering parameters: Three parameters control the three consecutive clustering steps of fpocket: (i) the maximum distance between Voronoi vertices for the 1st clustering step, (ii) the maximum distance between two cluster centroids for clustering step 2, and (iii) the maximum distance between two alpha sphere centres (Voronoi vertices) for the multiple linkage clustering step.

- Pocket pruning parameters: It is controlled by three parameters: (i) The minimum number of alpha spheres in a putative binding pocket, to prune too small clusters, (ii) the minimum ratio of apolar alpha spheres over the total number of spheres to prune too hydrophilic pockets – currently not in use.

On exit, fpocket will return different files containing information about the identified pockets. First, it will return a PDB file containing all atoms used for pocket detection from the input PDB file (ligands are discarded on input unless explicitely notified), supplemented by the positions of all alpha sphere centres (Voronoi vertices) retained after pocket detection. Voronoi vertice positions are added as HETATM in the PDB file. The residue name associated to these vertices is STP (for StTePoint). Residue numbers are given according to the pocket numbering and thus ranking. One can distinguish two types of Voronoi vertices (encoded by the atom type column of the PDB convention) in the PDB output: (i) APOL, for apolar vertices and (ii) POL, for polar vertices. Second it returns a file using the PQR convention that contains only the alpha sphere centres and radii. Again, residue numbers correspond to pocket ranks. Third, a file containing statistics about each pocket is provided. It lists different characteristics and scores of pockets identified on the surface of the protein. Fourth, scripts are provided, intending to render easier visualisation of putative binding pockets using PyMol or VMD. Finally, a per pocket series of files is also provided. For each pocket, a PDB file containing only the atoms defining the pocket and a PQR containing only the alpha spheres of the pocket are written.

Tpocket

Tpocket (Testing pockets) has been designed as a framework for the evaluation of the performance of the pocket detection algorithm and the accuracy of the implemented scoring function: Users trying to implement their own scoring functions can easily assess their performance using tpocket. The general workflow of this framework is presented on figure 1b. Generally one wants to assess a scoring function on a collection of PDB structures for which the binding site is known. In addition, it can be of importance to compare the performance of pocket detection for both apo and holo forms of the same protein. Tpocket can manage both constraints using an input list file, where each line should contain the information about one pair of related apo/holo structures: "path_to_the_apo_structure path_to_the_bound_structure name_of_the_ligand " and the name of the ligand is specified using the same 3 letter code (residue name) as in the PDB file. Note that when assessing the performance of fpocket using a set of apo/holo structures, the two forms should be superposed prior to the analysis.

The tpocket output is split up in two files. First, global performance for all available evaluation criteria described later is provided in a simple text file. Second, more detailed information about pocket detection is written in a separate text file for each structure, including the total number of pockets retained, all evaluation criteria implemented, the rank of the actual pocket detected by fpocket for each criteria, and some other values such as ligand and pocket volume evaluation, number of atoms in the pocket... Among other things, this file allows the identification of structures for which the fpocket detection failed (either because the pocket found has a low rank or was not found at all) for each evaluation criteria.

The fpocket prediction performance presented in this paper are based on tpocket results. Consequently, besides careful manual inspection of pocket evaluation results, they were further validated by an external evaluation script. A SVL (Scientific Vector Language) script was developed using the Moe Software from the CCG. This script evaluates fpocket performances based on fpocket output only. Tpocket and the SVL script gave both exactly the same result.

Dpocket

Dpocket (Describing pockets) is designed to organise descriptor collecting from a series of co-crystallized complexes. It accepts a list of structures to analyse using a file containing the information about one structure per line, on the form:

"path_to_the_structure name_of_the_ligand1"

For each structure, dpocket extracts several simple descriptors using atom, amino acid and alpha sphere information. Currently, the set of descriptors implemented is

related to (i) alpha spheres (number, polarity, density, ...) (ii) protein atoms (electronegativity, ...) (iii) residues (residue type occurrences, hydrophobicity, ...) (iv) volume. Additionally, some of these descriptors are normalised between 0 and 1 to allow comparison between pockets of different proteins. Although many of these descriptors are basic, users can easily implement more sophisticated analysis of pocket properties. Besides, the current scoring function shows impressive performance and is based on only 5 of these simple descriptors.

Dpocket provides three different output files. First, descriptors extracted from alpha spheres next to the ligand, are written in a separate text file. Second, descriptors for correctly identified binding pockets are extracted. Last, descriptors for other pockets found by fpocket are extracted in a separate text file. Detailed information on each descriptor used in the current version can be found in the full documentation.

Parameter optimisation

In order to determine optimal parameters for fpocket, a data set based on the protein test set used by An et al. in 2005 [16] for the evaluation of PocketFinder was used. The set described by An et al., composed of 5616 protein ligand complexes and 11510 apo forms is rather redundant, despite the fact that 5616 complexes are composed of the combination of 4711 unique proteins and 2175 unique ligands. The structural redundancy was eliminated allowing a maximum sequence identity of 50% between different proteins of this set. The PDB blastclust file, available on the PDB website was used for this purpose ftp://ftp.wwpdb.org/pub/pdb/derived_data/NR/blastclust/pdbS50bF.out. This first filter resulted in 307 proteins that we further validated by hand, in order to perform training on well defined binding pockets. Monomers and homomultimere containing more than one single binding pocket for the same ligand were removed. No particular filters were applied to the ligand type, as the druglike concept is still a matter of debates. During training, all hetero atoms were dropped from the PDB structure and pocket detection was performed not taking into account hydrogen atoms. Only structurally important HETATM recordings, like hemes, zinc etc. were kept in order to detect a "biologically" available binding pocket. A complete list of kept HETATM recordings is available in fpocket manual.

Currently, fpocket contains standard parameters determined by an semi combinatorial/empirical optimisation step using this training set. Basically, the fpocket parameters allow enough flexibility to obtain many small pockets as well as few very large pockets. During this optimisation, our goal was to clearly identify the pockets using two main pocket identification criteria (e.g. a good ligand coverage and a low distance value according to the PocketPicker distance criterion). Pockets found by the algorithm should be neither too small nor too large. To do so, it was intended to obtain a good relative overlap (e.g. size of the pocket found by fpocket/size of the actual pocket). Additionally, we attempted to minimise the number of pockets returned by the algorithm. The resulting fpocket standard parameter values are an alpha sphere minimum (resp. maximal) size of 3.0 (resp. 6.0) Å, a minimum connection distance 1 (resp. 2, 3) of 1.73 (resp. 2.5 and 4.0) Å, a minimum number of alpha spheres of 35.

Scoring function

Fpocket currently uses a simple 3 component PLS derived scoring function. This scoring function makes use of the ligand coverage as the dependant variable, and of the five following descriptors implemented in dpocket as independent variables: (i) the normalised number of alpha spheres, (ii) the normalised mean local hydrophobic density, (iii) the normalised proportion of apolar alpha sphere, (iv) the polarity score (sum of polarity over all amino acids involved in a given pocket using a binary scheme, e.g. 1 for polar, 0 for non polar) and (v) the alpha sphere density, defined as the mean value of all alpha sphere pair to pair distances in the binding pocket.

Note that the normalisation here means that the basic descriptor was scaled to a 0–1 range, so that for example the largest and the smallest pocket within a given protein would have a normalised number of alpha spheres of 1 and 0, respectively.

The model was trained using the dpocket output statistics run on the training dataset previously defined. No additional normalisation of descriptors (such as mean centring) was used as no difference was shown in terms of prediction accuracy.

Site identification assessment

In order to assess pocket prediction performance, one has to compare identified pockets to the real binding pocket. Different approaches exist in order to do so. Fpocket implements different methods to assess whether a binding pocket was found or not.

PocketPicker criterion (PPc): This is the criterion used in the PocketPicker [23] study. Here the geometric centre of the pocket is calculated. If the position of this centre is within 4 Å from any atom of the ligand, the binding site is considered correctly identified.

Mutual Overlap criterion (MOc): This criterion considers a pocket successfully identified if at least 50% of the ligand atoms lies within 3 Å of at least one alpha sphere, AND if at least 20% of the pocket alpha spheres lie within 3 Å of the ligand. In other words, the

first condition ensures that the ligand is at least half covered by the pocket, and the second condition allows the pocket to be quite large, but not too much as a significant proportion of probe still has to lie next to the ligand. Note that pockets larger than the effective region of interaction with the ligand have to be considered since several ligands may bind to different regions of the pocket (see Figure 2).

The MOc is introduced for two main reasons: (i) to further validate fpocket and see if it's performance remains acceptable using a rather different evaluation criterion and (ii) to address two issues related to the PPc.

Firstly, PPc does not ensure that a reasonable fraction (e.g. one half) of the ligand lies within the pocket identified. For example, a small cluster of probes (alpha spheres for fpocket) next to the ligand could be considered as a successful identification of the pocket even if none of the ligand atoms actually lies within the pocket volume. Secondly, large pockets generally cannot be considered as successfully identified using this criterion. Although it ensures that very large pocket (e.g. the whole protein), are considered as failure, we believe that this criterion is too restrictive, especially (i) when the ligand is small and/or not located at the centre of the pocket found, (ii) when the pocket is simply very large (large protein, multimer...) and (iii) when the pocket does not have a simple globular form.

Figure 3 illustrates differences between the two criteria. Here, pockets are considered as successfully (1esa PDB entry) and unsuccessfully (1w1p PDB entry) identified by PPc, respectively. However, for the 1esa case, one cannot consider the pocket as successfully identified, as only a small part of the ligand is covered by the pocket; the MOc considers this case as a failure since less than 50% of the ligand is covered by the pocket. For 1w1p, PPc fails, mainly because the ligand is not located at the centre of the pocket, and because the pocket is rather large; the MOc considers this case as a successful one, as the ligand is covered at 100% and 25% of the alpha spheres lie next to the ligand.

Results

Evaluation of pocket prediction accuracy

Table 1 presents fpocket performance on 3 different data sets. The first one consists in a collection of 48 proteins [23] already used in a previous study for which results of several methods on the bound and unbound conformations are reported. In order to keep the comparison valid, we haven't modified this dataset, although we identified several cases of multiple binding sites that should be removed in a rank-based evaluation. The second one was derived from a contribution by Alan C Cheng & al. [39]. They used a set of 63 structure representing 27 pharmaceutical targets, including 23 targets with marketed drugs or drugs in Phase II or above. We have selected randomly one protein-ligand complex for each of these targets to avoid redundancy, and the same filters as those used or the parameter optimization set were applied, resulting in a set of 20 pdb files. Finally, the recently defined Astex diverse set [40] was used. This dataset consists of 85 diverse high resolution protein-ligand crystal structures retrieved from the PDB using newly developed analysis and classification techniques. This last dataset has been built using the following filters: (i) the ligand is drug-like; 23 of the ligands are approved drugs and 6 are currently in clinical trials (ii) no particular target is represented more than once (iii) the proteins are all drug discovery or agrochemical targets (iv) only high quality structures are included for which the ligand electron density supports the entire ligand binding mode (v) no structures are included where the ligand is in contact with protein atoms of crystal symmetric units After applying our filtering procedure, 82 proteins were kept. For sake of comparison, results obtained using the Pocket Picker criterion (PPc) are first discussed. From the proteins in complex with the ligand, fpocket correctly identifies 83% (resp. 92%) of the actual pockets within the top 1 and top 3 ranked pockets, a performance better than other approaches. From the unbound conformations of these proteins, the corresponding results are of 69% and 94%, respectively. At rank 1, similarly to other approaches, fpocket performance decreases, but remains however better than all methods evaluated on this dataset, except LIGSITE (csc) that shows a slightly better performance (2%) and Pocket-Picker for which fpocket reaches similar score. At rank 3, fpocket outperforms by far all other approaches except possibly LIGSITE (csc) for which no result at this rank is available. This would indicate that fpocket's pocket detection is particularly efficient, and that further filtering on pocket drugability (for instance) could be used to re-rank the top 3 pockets. In order to test the robustness of fpocket depending on the dataset, we also present the results of fpocket and Pocket Picker on two other sets. At rank 1, we observe for fpocket scores of 75 and 67% on the Cheng and Astex diverse sets, respectively. Fpocket scores better than Pocket Picker by 5 and 8% respectively. In addition, again one could note that the fpocket performance at rank 3 remains by far higher.

In Table 1 are also listed the fpocket performances using the mutual overlap criterion (MOc) introduced in this paper. Compared to the PPc, no significant differences are observed in terms of performance measures for the Pocket Picker set, but slightly smaller (resp. better) performance measures on the Cheng (resp. Astex diverse) set. However, on average, the performance at rank 3 remains more stable, close to 90% using the MOc. Looking more in detail,

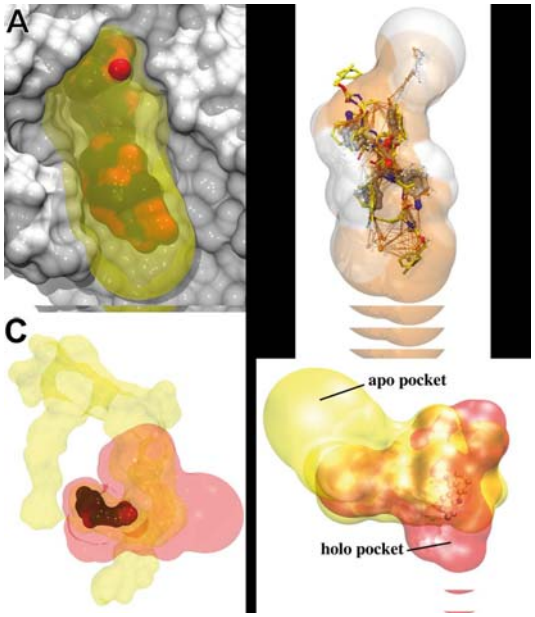


Figure 2
Examples of pocket detection using fpocket. **top left**: Rank 1 pocket on the alpha amylase (7TAA). Acarbose in surface/coloured/opaque representation, the binding site is represented as yellow transparent hull. Alpha sphere centres are depicted as small red points. **top right**: Rank 1 pocket of the HIV1 Protease DMP450 complex (PDB Code: 1DMP). DMP450 is depicted in grey CPK representation and the binding pocket as transparent hull. Superposed are other known inhibitors (yellow) binding in the same pocket (PDB Codes: 1Z1H, 2UY0, 2P3B). Alpha sphere centres are depicted as small interconnected spheres. Alpha spheres and the pocket are coloured according to polar (orange) and apolar (white) character. **bottom left**: Cyclooxygenase-2 indomethacin binding site: (red) pocket identified by fpocket,(yellow) pocket identified by PocketPicker. **bottom right**: Acetylcholinesterase rank 1 predicted binding pocket by fpocket. Red: pocket of the holo structure with tacrine (1ACJ), yellow: pocket of apo structure (1QIF). Pockets are represented as a hull resulting from the union of the alpha spheres.

the 5% difference of observed for the Cheng set only represent one protein, due to the low number of structures in this set. The Astex diverse set contains 6 proteins for which the MOc and PPc disagree, and for all of them, the MOc detects the pocket correctly while the PPc does not. Pocket size seems to be the major issue. In the Astex dataset, the mean number of atoms per pocket is 91 (defined as all unique atom contacted by alpha spheres within the pocket). For the 6 cases mentioned previously, this number ranges between 116 and 281. This illustrates a better behaviour of the MOc on larger pockets for which PPc seems unadapted – see methods.

Examples of successful identification of binding sites

Figure 2a shows the successful identification (rank 1) of the acarbose binding pocket on alpha amylase (PDB code 7taa). Acarbose is represented in coloured surface and the fpocket identified binding pocket as transparent hull around the ligand. This rather long and large pocket has a buried and a more solvent exposed part. Despite this heterogeneity within the whole binding pocket, fpocket identifies the whole pocket with a reasonable pocket volume around the ligand.

On figure 2b another interesting feature about fpocket is shown. Here the binding pocket of HIV1 protease is depicted in complex with the Dupont Merck inhibitor DMP450 (PDB code 1dmp). For representative reasons the protein structure was omitted and only the surface of the pocket is shown (alpha sphere surface) with the embedded ligand. The small interconnected spheres are the alpha sphere centres. Orange alpha spheres are polar alpha spheres, white alpha spheres are apolar. The same colour code was used for the colouring of the pocket surface. Here, one can notice that the positions of alpha sphere centres follow surprisingly well the topology of the ligand (grey). Note, however, that this is not a general property of Voronoi vertices. Next, physicochemical properties of the ligand are reflected by the sourrounding binding pocket. The pocket identified by fpocket seems far longer than the actual binding position of the ligand. However other drug like molecules (yellow) are known to make interactions also with residues situated on the edge of the pocket (top and bottom here).

These examples show that fpocket is able to detect solvent exposed and very buried binding sites, that bind ligands of a very different nature (oligosaccharide, drug)

Last an example of cyclooxygenase-2 indomethacin complex (PDB code 4cox) is depicted on figure 2c. The binding pocket identified using PocketPicker is represented as yellow halo. As red halo one can find the fpocket identified binding pocket. Both binding pockets include the actual space of the pocket occupied by the ligand, but the PocketPicker yields a far bigger pocket, including sourrounding channels.

Examples of unsuccessful identification of binding sites

Figure 2d depicts one example of a binding site that was not correctly identified according to the PPc (see methods). These structures are part of the PocketPicker data set. Here the acetylcholinesterase active site gorge was successfully identified and ranked on the holo form (PDB code: 1acj). The pocket is represented as red envelope. The same pocket on the apo form (PDB code: 1qif) depicted in yel-

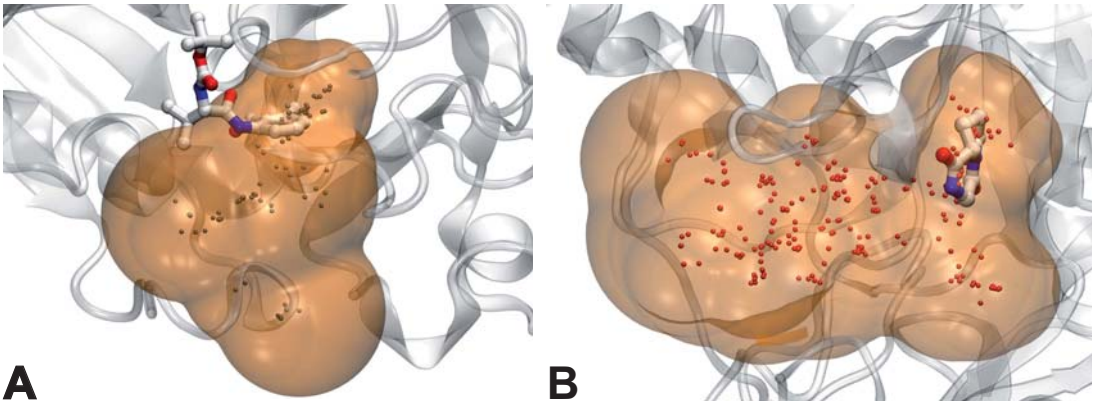


Figure 3
Pocket detection limits. **Left**: Example of PDB entry 1esa. A large part of the ligand is outside the pocket detected by fpocket. Despite this fact, a criterion such as the PocketPicker criterion would accept the pocket as successfully identified, and the Mutual Overlap criterion not. **Right**: Example of PDB entry 1w1p. The identified pocket is large compared to the ligand. Its centre of mass is too far from any atom of the ligand for the Pocket Picker criterion to accept it as successfully identified. Ligands are represented using a ball and sticks representation. Alpha sphere centres are represented as small spheres, and their envelope is depicted in brown.

low in figure 3a shows a completely different shape compared to the holo form. This is due to the fact, that the binding pocket is very buried and upon closure of the binding site entry a longer binding pocket was identified. According to the PocketPicker criterion fpocket did not identify well the pocket in the apo form, although the identified binding pocket overlaps nearly completely the previously identified holo pocket. This example shows the limits of the criterion used by PocketPicker to distinguish correctly identified binding sites from others. The MOc overlap criterion presented here and used in similar ways in other studies shows better accordance to visual results than the simple distance criterion used by PocketPicker.

Table 1: Fpocket performance

Dataset	Algorithm	Rank 1		Rank 3	
		unbound	bound	unbound	bound
Pocket Picker					
	Fpocket	69 (67)	83 (85)	94 (92)	92 (92)
	PocketPicker	69	72	85	85
	LIGSITE(CS)	60	69	77	87
	LIGSITE	58	69	75	87
	CAST	58	67	75	83
	PASS	60	63	71	81
	SURFNET	52	54	75	78
	LIGSITE(CSC)	71	79	-	-
Cheng et al.					
	Fpocket	-	75 (70)	-	95 (90)
	PocketPicker	-	70	-	80
Astex Diverse set					
	Fpocket	-	67 (73)	-	82 (88)
	PocketPicker	-	59	-	67

Comparison of results obtained for fpocket and other approaches. For sake of comparison, scores are reported using the PPc, and we present scores at rank 1 and 3 (true pocket in the top 3 pockets proposed by fpocket). For the Pocket Picker dataset, results are taken from [23] for all but fpocket. For fpocket, numbers within parentheses correspond to scores obtained using the MOc.

Computational performance

The algorithm was assessed on a Intel Celeron M 1.6 Ghz, 1 Gb RAM architecture and performed roughly one structure per less than one to three seconds, depending on the size of the structure. For the sake of completeness, performance of LigSite and PASS was compared on the same structures. LigSite performed pocket detection on one structure in 5 seconds, PASS in 4 to 5 seconds. Thus, fpocket appears particularly well suited for large scale evaluations and is situated among the fastest algorithms in the field.

PocketPicker performs roughly one structure in several hours of calculation depending on the size of the structure.

Conclusion

We have introduced fpocket, a new open source pocket identification platform. Compared to other approaches of the field, Fpocket performs well on state of the art data sets. From the complexed protein conformations, fpocket reaches the best performance at rank 1. On the ligand free structures, similarly to other approaches, fpocket performance drops at rank 1, but is much better at rank 3, outperforming other approaches by more than 9%, opening the door to further pocket drugability filtering approaches. Interestingly, fpocket is among the fastest algorithms in the field. This makes fpocket particularly well suited for high throughput pocket detection and construction of cavity databases. Next, fpocket comes with its underlying programs, tpocket and dpocket, providing powerful research tools for a large scale assessment of own pocket scoring functions and properties of binding pockets, respectively. Its open source character provides a useful contribution to the scientific community willing to further develop and research in the pocket identification and specific molecular binding field.

Availability and requirements

Fpocket source code (Linux) is freely available under the GNU GPL license at http://fpocket.sourceforge.net. The required Qhull package is shipped and compiled together with fpocket in the official fpocket release.

Authors' contributions

VLG and PS have equally contributed to fpocket development. PT has initiated and supervised fpocket development.

Acknowledgements

The authors thank Xavier Barril Alonso for helpful discussions on PDB input and descriptor development for fpocket.

References

1. Manly CJ, Chandrasekhar J, Ochterski JW, Hammer JD, Warfield BB: **Strategies and tactics for optimizing the Hit-to-Lead process and beyond-A computational chemistry perspective.** *Drug Discov Today* 2008, **13(3–4):**99-109.
2. Villoutreix BO, Bastard K, Sperandio O, Fahraeus R, Poyet JL, Calvo F, Déprez B, Miteva MA: **In silico-in vitro screening of protein-protein interactions: towards the next generation of therapeutics.** *Curr Pharm Biotechnol* 2008, **9(2):**103-22.
3. Totrov M, Abagyan R: **Flexible ligand docking to multiple receptor conformations: a practical alternative.** *Curr Opin Struct Biol* 2008, **18(2):**178-84.
4. Levitt DG, Banaszak LJ: **POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids.** *J Mol Graph* 1992, **10(4):**229-34.
5. Delaney JS: **Finding and filling protein cavities using cellular logic operations.** *J Mol Graph* 1992, **10(3):**174-7.
6. Del Carpio CA, Takahashi Y, Sasaki S: **A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I). Search for pocket regions.** *J Mol Graph* 1993, **11:**23-9. 42
7. Kleywegt GJ, Jones TA: **Detection, delineation, measurement and display of cavities in macromolecular structures.** *Acta Crystallogr D Biol Crystallogr* 1994, **50(Pt 2):**178-85.
8. Laskowski RA: **SURFNET: a programm for visualizing molecular surfaces, cavities and intermolecular interactions.** *J Mol Graph* 1995, **13:**323-330. 307–308
9. Masuya M, Doi J: **Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations.** *J Mol Graph* 1995, **13(6):**331-6.
10. Peters KP, Fauck J, Frommel C: **The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria.** *J Mol Biol* 1996, **256:**201-213.
11. Hendlich M, Rippmann F, Barnickel G: **LIGSITE: automatic and efficient detection of potential small-molecule binding sites in proteins.** *J Mol Graph Model* 1997, **15:**359-363. 389
12. Ruppert J, Welch W, Jain AN: **Automatic identification and representation of protein bvinding sites for molecular docking.** *Protein Sci* 1997, **6:**524-533.
13. Liang J, Edelsbrunner H, Woodward C: **Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.** *Protein Sci* 1998, **7:**1884-1897.
14. Brady GP Jr, Stouten PF: **Fast prediction and visualization of protein binding pockets with PASS.** *J Comput Aided Mol Des* 2000, **14:**383-401.
15. Venkatachalam CM, Jiang X, Oldfield T, Waldman M: **LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites.** *J Mol Graph Model* 2003, **21(4):**289-307.
16. An J, Totrov M, Abagyan R: **Pocketome via comprehensive identification and classification of ligand binding envelopes.** *Mol Cell Proteomics* 2005, **4:**752-761.
17. Laurie A, Jackson R: **Q-SiteFinder: an energy-based method for the prediction of protein-lidang binding sites.** *Bioinformatics* 2005, **21:**1908-1916.
18. Nayal M, Honig B: **On the nature of Cavities on protein surfaces: application to the identification of drug-binding sites.** *Proteins: Struc Func Bioinform* 2006, **6:**892-906.
19. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.** *BMC Struct Biol* 2006, **2006(6):**19.
20. Coleman RG, Sharp KA: **Travel depth, a new shape descriptor for macromolecules: application to ligand binding.** *J Mol Biol* 2006, **362(3):**441-58.
21. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM: **A method for localizing ligand binding pockets in protein structures.** *Proteins* 2006, **62(2):**479-488.
22. Bock ME, Garutti C, Guerra C: **Effective labeling of molecular surface points for cavity detection and location of putative binding sites.** *Comput Syst Bioinformatics Conf* 2007, **6:**263-74.
23. Weisel M, Proschak E, Schneider G: **PocketPicker: analysis of ligand binding-sites with shape descriptors.** *Chem Cent J* 2007, **1(7):**1-17.
24. Kawabata T, Go N: **Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites.** *Proteins* 2007, **68(2):**516-29.
25. Xie L, Bourne PE: **A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites.** *BMC Bioinformatics* 2007, **22(Suppl 4):**S9.

26. Kim D, Cho CH, Cho Y, Ryu J, Bhak J, Kim DS: **Pocket extraction on proteins via the Voronoi diagram of spheres.** *J Mol Graph Model* 2008, **26(7):**1104-12.
27. An JTMAR: **Comprehensive identification of "druggable" protein ligand binding sites.** *Genome Inform* 2004, **15(2):**31-41.
28. Zhong S, MacKerell ADJ: **Binding response: a descriptor for selecting ligand binding site on protein surfaces.** *J Chem Inf Model* 2007, **47(6):**2303-2315.
29. McGovern SL, Shoichet BK: **Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes.** *J Med Chem* **46(14):**2895-2907.
30. Bhinge A, Chakrabarti P, Uthanumallian K, Bajaj K, Chakraborty K, Varadarajan R: **Accurate detection of protein:ligand binding sites using molecular dynamics simulations.** *Structure* 2004, **12(11):**1989-1999.
31. Yang AY, Källblad P, Mancera RL: **Molecular modelling prediction of ligand binding site flexibility.** *J Comput Aided Mol Des* 2004, **18(4):**235-250.
32. Murga LF, Ondrechen MJ, Ringe D: **Prediction of interaction sites from apo 3D structures when the holo conformation is different.** *Proteins* 2008, **72(3):**980-92.
33. **The Chemical Computing Group** [http://www.chemcomp.com/]
34. **Q-SiteFinder Ligand Binding Site Prediction** [http://www.modelling.leeds.ac.uk/qsitefinder/]
35. **Pocket-Finder Pocket Detection** [http://www.modelling.leeds.ac.uk/pocketfinder/]
36. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J: **CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues.** *Nucleic Acids Res* 2006, **1(34 Web Server):**W116-8.
37. Goto J, Kataoka R, Muta H, Hirayama N: **ASEDock-docking based on alpha spheres and excluded volumes.** *J Chem Inf Model* 2008, **48(3):**583-90.
38. **The Quickhull algorithm for convex hulls** [http://www.qhull.org]
39. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES: **Structure-based maximal affinity model predicts small-molecule druggability.** *Nat Biotechnol* 2007, **25:**71-5.
40. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW: **Diverse, high-quality test set for the validation of protein-ligand docking performance.** *J Med Chem* 2007, **50(4):**726-41.

Druggability Prediction

Daniel Alvarez-Garcia, Jesus Seco, Peter Schmidtke and Xavier Barril

Protein Ligand interactions. Edited by H. Gohlke. Series "Methods and Principles in Medicinal Chemistry". Eds. R. Mannhold, H. Kubinyi & G. Folkers. Wiley-VCH. 2011. Accepted

Druggability Prediction

Daniel Alvarez-Garcia,[a] *Jesus Seco,*[a] ***Peter Schmidtke*** [a] *and Xavier Barril*[a,b]

[a] Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona Av. Joan XXIII s/n, 08028 Barcelona (Spain) and Institut de Biomedicina de la Universitat de Barcelona (IBUB). [b] Catalan Institution for Research and Advanced Studies (ICREA)

Protein Ligand interactions. Edited by H. Gohlke. Series "Methods and Principles in Medicinal Chemistry". Eds. R. Mannhold, H. Kubinyi & G. Folkers. Wiley-VCH. 2011. Accepted

1 INTRODUCTION

Modern drug discovery starts with the selection of a target biological component, usually a protein, whose activity is associated with the condition for which a therapy is sought. Target selection involves a thorough revision of the biological and pathophysiological data, sometimes involving a target validation stage during which the therapeutic potential of the target is assessed [1]. By nature, the target selection and target validation processes focus on molecular biology aspects, but it must also consider the ability of drug-like molecules to bind and to alter the biological activity of the target (target druggability). Most known drugs act on a handful of target types (Fig. 1) and protein classification is often taken as an indication of druggability. However, such an approximation may lead to the exclusion of valid targets (e.g. certain protein-protein interactions) or to the prioritization of targets that turn out to be intractable; for instance, 50% of drug discovery programs targeting enzymes at GSK failed to produce viable leads [2]. The latter situation results in a major waste of resources, whereas the former leads to a loss of opportunities. Understanding and predicting when a target will be druggable is therefore of the utmost importance in pharmaceutical research. In the next two sections, the determinants of target druggability are discussed. We then proceed to present experimental and computational means to predict or assess this property. This is followed by the presentation of a test case that enables us to illustrate some of the concepts introduced. Finally, the main points are reviewed and a perspective is offered.

2 DRUGGABILITY: LIGAND PROPERTIES

The physical and chemical properties of small organic molecules determine, to a large extend, their fate in a biological system. In the case of drugs, an adequate administration, distribution, metabolization and excretion (ADME) profile is essential to reach the site of action. Considering that oral administration is the most desirable route (and the most difficult to achieve) the properties of oral drugs have been thoroughly investigated and they set the gold standard for drug candidates. As initially proposed by Lipinski and co-workers, drug absorption and permeability are more likely for molecules with molecular weight below 500Da, LogP (a measure of lipophilicity) below 5, number of hydrogen bond acceptors less than 10 and number of hydrogen bond donors less than 5 [4]. The so-called rule of five has gained widespread acceptance and is supported by more recent and detailed studies that reinforce the idea that increased molecular weight or excessive lipophilicity are a handicap for oral drug candidates [5-8]. As drugs must achieve tight binding (often below 10 nM), a high level of complementarity must exist between the ligand and the receptor. This usually requires that the latter wraps around the former, thus increasing the contact area. Considering the rules of molecular recognition, the physical and chemical properties of druggable binding sites must mirror those of drugs [9]. In summary, to be druggable, a binding site must consist of a surface that grants maximal shape complementarity (i.e. concave) and it must present a balance of polar and apolar features matching those of a drug-like ligand.

3 DRUGGABILITY: LIGAND BINDING

Most biomacromolecules have a tendency to interact with other organic molecules (proteins, nucleic acids, metabolites, etc.), but this property is unevenly distributed over the protein surface. Surface patches with a larger interaction potential are known as ‘hot spots’, a concept originating from alanine scanning experiments that probed the interface between protein-protein complexes. These studies revealed that most of the binding free energy is contributed by a few residues only [10,11]. The O-ring theory, initially introduced by Bogan & Thorn, suggests that hot-spot residues are easily desolvated because the local environment induces solvent-exclusion [12-14]. Although structurally different, ligand binding sites display the same behaviour: some regions in the binding area interact very favourably with particular functional groups, while the rest may provide the right shape and solvent exclusion capacity [15]. The presence of hot spots is, therefore, necessary for binding to occur, but what are the distinct properties of druggable

binding sites? This has been investigated in parallel with the development of druggability prediction methods.

As expected from the properties of drug-like ligands (see above), closed and lipophilic binding sites are more likely to be druggable. This is supported by the usefulness of related parameters to obtain predictive models following inductive [16,17] or deductive [18] reasoning. Binding site curvature relates to the necessity to maximize the contact surface area between the ligand and the protein, while the positive correlation of apolar surface area with druggability would presumably suggest that binding potency is entirely due to hydrophobic interactions. This was justified on the basis that electrostatic interaction and desolvation energies act in opposition, resulting in an overall negligible contribution [16,18]. However, this contradicts the empirical observation that polar interactions often constitute anchoring points, featuring predominantly in pharmacophoric models of binding sites. In fact, the contribution of polar interactions is context dependent, and a single hydrogen bond can contribute as much as 1.8 kcal/mol [19], comparable to the hydrophobic gain provided by the side-chain of a Val residue [20]. The increased proportion of apolar surface area in druggable binding sites compared to non-druggable (70% vs. 50%) can, actually, be reconciled with the importance of polar interactions: polar atoms in druggable binding sites are less solvent exposed and have a predominantly hydrophobic environment, resulting in lower dielectric environment that potentiates electrostatic interactions. This effect has been quantified in proteins, demonstrating that hydrogen bonds can be up to 1.2 kcal/mol stronger in hydrophobic environments [21]. This clearly indicates that – beyond the obvious gain in hydrophobic potential – a decrease in the polar surface ratio can have the paradoxical effect of increasing the hydrogen bonding potential of the binding site.

An intriguing property of druggable binding sites is that, in spite of being mostly buried, polar atoms protrude more from the cavity surface than apolar atoms [17]. In such disposition they are readily available to interact with incoming ligands, providing anchoring or selectivity points. It has been suggested that this type of environment can also transform polar atoms into kinetic traps. The molecular mechanism consists in a simple decoupling of the ligand/water exchange processes due to the steric impediments imposed by the local environment. In such circumstances, protein-ligand hydrogen bonds must start to break before a water molecule can mediate the process (and vice-versa), thus penalizing the exchange process and slowing down diffusion rates (Schmidtke et al., submitted). This might stabilize transient encounter complexes, facilitating their mutual recognition [22] and, once formed, lock the binding mode of the interacting molecules.

4 DRUGGABILITY PREDICTION BY PROTEIN CLASS

An inspection of Figure 1 suggests that enzymes and receptors are druggable targets, while other protein classes are difficult to undruggable. This is generally a good assumption because these protein classes have evolved to interact with small organic molecules (substrates, hormones, neurotransmitters, etc.), which means that drugs can compete in equal terms and achieve high affinity for the binding site. Depending on the type of natural substrate, their difficulty as targets will also vary: proteins binding *bona fide* small molecules (e.g. class A GPCRs, kinases) are more druggable than those binding non-druglike ligands, such as peptides (e.g. classes B and C GPCRs, proteases). The influential paper by Hopkins and Groom and other works estimating the size of the so-called "druggable genome" [3,23] relied on protein domain annotations (e.g. from PFAM [24]) to predict the number of proteins containing domains experimentally known to be targeted by small-molecule drugs. A support vector machine method has also been developed to predict protein druggability based on amino acid sequence independent of sequence similarity [25]. However, the financial disclaimer "past performance is no guarantee of future results" also applies here and, while those approaches may be valid at a statistical level, the chances of success when selecting a particular target becomes a probability game. Another weakness of such approaches is that they potentially ignore some mechanisms of action that are not linked to a particular sequence or domain topology. Allosteric modulation [26] and protein-protein inhibitors [27,28] are two examples of target types that, although each individual structure has a low probability of being druggable, their ubiquity and abundance (current estimates for the human interactome are 130,000 protein-protein pairs [29]) warrants many new therapeutic opportunities.

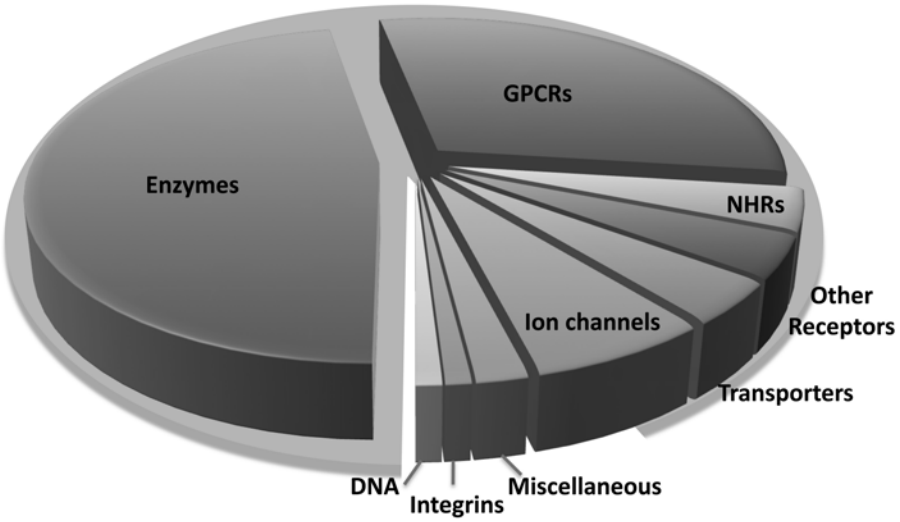


Figure 1. Targets of small-molecule drugs classified by function or protein family. Data taken from reference [3].

5 DRUGGABILITY PREDICTIONS: EXPERIMENTAL METHODS

5.1 High-Throughput Screening

In theory, the most rigorous form of assessing druggability is to test the ability of drug-like compounds to modify the biological activity of the target of interest. In this regard, examining the rate of success of high-throughput screening (HTS) results is very informative. In 2006, Macarron published a retrospective analysis of HTS campaigns at GlaxoSmithKline, grouping the results by target families. The analysis is particularly interesting because success was defined as the ability to produce a confirmed hit (i.e. activity in a biologically relevant assay with a tractable chemical structure and an initial indication of SAR such that a chemical optimization effort can begin). As expected, some target families offer very good results (e.g. a lead could be identified from HTS in >70% of Nuclear Hormone Receptors and Ion channels), whereas the success rate was only 33% for targets not belonging to the main classes [2]. Gupta et al., from AstraZeneca also published a retrospective analysis of HTS on 22 enzymes (identified by

function), but in this case the hit rate (i.e. percentage of compounds with read-outs about a certain threshold) was reported as a measure of success [30]. The values reported range from 0.06% to 3.85% for a common collection of 37,275 compounds. As pointed out by Macarron, one of the limitations of these retrospective analyses is that it is not possible to know if failure happens because the target is undruggable or because the collection of compounds tested does not cover the adequate chemical space: a 30% of al targets that failed when tested on a subset of the historical collections turned out to be tractable when tested against the unified GSK collection [2]. Considering the vastness of the drug-like molecular space (estimated at 10^{20}-10^{24} synthetically accessible compounds [31]), this is an important issue and suggests that success with novel target types is partly limited by the composition of current historical collections. It also raises questions about the usefulness of hit rates as druggability predictions. One should also be aware of the limitations of the specific assay, for instance, a binding assay may not be the most suitable to identify allosteric modulators.

5.2 Fragment Screening

Fragment screening was initially described in 1996 [32], adopting new detection methods and becoming an extremely popular hit identification strategy in the 2000s [33-36]. Its main advantage is the superior ability to detect binders because it explores much simpler compounds than HTS [37]. Considering that the number of possible chemical compounds grows as a quadratic function with the number of atoms [38], even if the number of compounds tested is usually 3 orders of magnitude smaller than HTS, it can in fact explore a much larger proportion of the corresponding chemical space. In consequence, the fragment screening hit rates may be more informative about the druggability of a given protein than those coming from HTS. Abbot and Vernalis have published data for 23 and 12 targets, respectively [16,39], and in both cases there is a good correlation between poor hit rates and the difficulty to obtain high affinity ligands. This is a strong indication that fragment screening may be a suitable method to detect good binding sites for small molecules. Once the necessary infrastructure and know-how is in place, the cost of fragment screening and the time needed to set up the experiment is much lower than the corresponding HTS assay, so carrying out a fragment screening experiment before launching a full drug discovery project may be a wise and feasible approach for small and large pharmaceutical companies. One potential limitation of this approach is that it is difficult to predict the drug-likeness of future ligands based on the chemical structure of the fragment hits. In other words, the method seems adequate to detect targets that do not offer binding opportunities, but does not warrant that hits can be developed into drugs.

5.3 Multiple Solvent Crystallographic Screening

Before the fragment screening era, it was detected that organic solvents have a large propensity to interact with binding sites of proteins both in solution and in crystals [40,41]. This raised the possibility of using 'solvent-mapping' to detect and characterize binding sites, something that has been achieved for a few systems [40,42,43]. In perspective, this can be seen as an extreme form of fragment screening: as the ligands tested are smaller, they are more likely to bind and fewer compounds need to be tested, but more sensitive methods are needed to detect binding. The detection method is precisely the limitation of this approach: few proteins form crystals sufficiently stable to withstand the high concentrations of organic molecules necessary to carry out Multiple Solvent Crystallographic Screening. It is, however, conceivable that current methods in fragment screening could be adapted to test simpler and weaker ligands with the specific aim of predicting druggability.

6 DRUGGABILITY PREDICTIONS: COMPUTATIONAL METHODS

6.1 Cavity Detection Algorithms

Due to the shape complementarity requisite, the binding site of ligands correspond to protein surfaces with inward curvature. Deep pockets are generally assumed to play a functional role and, in consequence, cavity detection algorithms have long been used to predict ligand binding sites. A large range of computer programs have been developed to identify pockets and to predict their likelihood to act as ligand binding sites (reviewed in [44]). The algorithms can roughly be classified into geometric or energetic approaches. In the first class –which is the most common– the protein shape is directly probed to detect void spaces surrounded by protein atoms. In the second approach the interaction energy of chemical probes (ranging from a simple sphere with van der Waals parameters to a diverse set of chemical fragments with van der Waals, hydrogen bonding and electrostatic potentials) is mapped on the three-dimensional space of the protein and ligand binding sites are identified on the basis of interaction energy profiles.

The main objective of those programs is to distinguish the true ligand binding site from the rest of cavities in a protein structure. As ligand binding sites often coincide with the largest protein pocket [15,45], size alone is a good predictor but most methods use a combination of parameters to rank the pockets, which in some cases also include information on residue

conservation. Success rates for the most recently published methods are close to 70% for the highest ranked pocket and 90% when the top 3 pockets are considered [46-48].

Achieving a representation of pockets that matches the space occupied by the ligands in an automated manner is far from trivial, because the ligand binding site is usually part of a larger network of pockets on the protein surface. However, cavity detection algorithms have a long history and have reached a fair level of maturity. At the same time, these programs are evolving to incorporate new functionalities that can be extremely useful in drug design [44]. These include consideration of pocket flexibility, pocket comparison algorithms and pocket druggability, which is discussed in the next section.

6.2 Empirical models

The first druggability prediction methods were developed at Abbott Laboratories [16] and Pfizer [18] to fulfil an unmet need in the pharmaceutical research industry. These and a number of more recently published methods build on cavity detection algorithms to extract pocket surface descriptors for druggability predictions. However, they differ in two main points from their parent methods: 1) pockets are compared not only within, but also across protein structures; and 2) instead of distinguishing binding sites from non-binding sites, their goal is to predict the likelihood that the pocket displays high affinity for drug-like ligands. Naturally, they require a completely new parameterization based on a distinct training set. In fact, obtaining a sufficiently large set of binding sites encompassing a wide range of druggability scores has been one of the main factors limiting progress in the field. It should also be noted that binding site druggability is a complex and somewhat fuzzy concept that can be defined in more than one way. As the predictions will be –at most– as good as the dataset on which the method has been trained, attention should be paid to the precise definition of druggability and to the composition of the training set. For this reason, here we focus only on published approaches that use a manually curated training set.

6.2.1 Training sets

The first druggability prediction method was trained to reproduce NMR fragment screening hit rates. The dataset consisted of 28 binding sites on 23 different proteins, on which 10,000 compounds were tested. Using heteronuclear NMR, perturbations anywhere on the protein can be detected and ligands with K_D values as high as 5 mM can be identified [32]. The physicochemical properties of the screening library conform to the definition of fragments (average molecular weight of 220 and an average cLogP of 1.5) and, being tested at high

concentrations (0.5 – 1.0 mM), they are highly soluble. The hit rates –ranging from 0.01% to 0.94%– were used as a measure of druggability [16]. As demonstrated in the paper, high correlation is observed between the experimental NMR hit rate and the ability to identify high-affinity (K_D < 300 nM) ligands. In line with this approach, researchers at AstraZeneca have used the HTS hit-rates as a measure of druggability. Using a set of 22 undisclosed targets, they obtain predictive models [30]. However, this definition of druggability presents two main limitations:

1. Ligand drug-likeness implicitly derives from the composition of the screening library, but its physicochemical properties can be very different from typical drugs, particularly in the case of fragments.
2. A major practical bottleneck is that screening data is proprietary, expensive to obtain and rarely made publically available. Additionally, extension of published datasets would require using the same screening library and methodology, limiting its transferability across organizations.

In 2007 Cheng and co-workers presented an alternative view of druggability, defined as the maximal affinity that a drug-like ligand (ideally an orally bioavailable compound) can achieve for a binding pocket [18]. This definition also presents some limitations, such as the fact that the druglikeness of a compound is sometimes difficult to assess or that the classification of a target may change over time. Obtaining good quality data can also be difficult, particularly when it comes to undruggable binding sites, because they can only be classified as such after substantial research efforts have been invested and negative data is often not published. However, the definition is useful in decision-making, because it can distinguish between targets that are likely to have a successful outcome (i.e. deliver an orally bioavailable lead) and those that are more likely to prove very challenging and may require other approaches (e.g. a pro-drug strategy). Subsequent druggability prediction methods have mostly adhered to this definition.

In order to facilitate further developments and to establish a benchmark that could be used in prediction performance, the initial set of 27 targets presented by Cheng et al. was extended by Schmidtke and Barril [17] with 1070 structures representing 70 different targets. The set was obtained crossing a list of oral drugs with information from the PDB [49] and the DrugBank [50], followed by visual inspection. The unified catalogue is publically available as the Druggable Cavity Directory (http://fpocket.sourceforge.net/dcd), a resource that can also be used to extend the dataset or to reassess target classification in a collaborative manner.

6.2.2 Applicability and prediction performance

Some of the published druggability prediction methods are difficult to apply because they used a combination of algorithms that included commercial and proprietary software that any potential user would have to reimplement [16,18,30]. An additional limitation of those methods is that cavity definition may also involve a manual procedure, which precludes their applicability in an unsupervised and high-throughput manner. Fortunately, more recent contributions can be used out of the box. Particularly noticeable in this regard are SiteMap, from Schrödinger, which includes a druggability score trained on Cheng's dataset [18] and the open source program fpocket [46], which provides a druggability score trained and tested on the afore-mentioned extended druggability dataset. Both can be used in an unsupervised manner and applied to large collection of structures to screen for druggable cavities, delivering similar performance[17]. For such applications, computational performance is also an important consideration and the Voronoi tessellation shape-based algorithm in fpocket is clearly superior to the grid-based interaction energy algorithm in SiteMap (2-4 seconds compared to several minutes). As different crystallographic structures of the same binding site may correspond to different conformations, the reproducibility of the druggability score must be assessed. With the exception of closed and rigid cavities, predictions may be substantially different due to changes in the properties of the cavity or to variability introduced by the automated cavity definition algorithm. However, both the mean values and the values of the top scoring cavities are clearly different between druggable and non-druggable cavities, suggesting that confidence in druggability predictions may increase when multiple structures are considered [17].

6.3 Physical chemistry predictions

Computational methods based on the principles of physical chemistry can be used to predict the interaction free energy between a ligand and a protein binding site. As this property is intimately linked to the druggability concept, molecular simulations offer an alternative to empirical approaches. The main difficulty in predicting binding free energies is that they are the end result of multiple terms of large and opposing magnitude. Consequently, accurate predictions are computationally very demanding and extremely hard to achieve[51]. The concept has nevertheless been used successfully in energy-based binding site detection methods, which rely on extremely crude but very fast approximations [44]. With increasingly rigorous approaches, it is theoretically possible to carry out the *in silico* equivalent of experimental druggability prediction methods. For instance, Huang and Jacobson have demonstrated that hit rates in docking-based virtual screening experiments correlate with the experimental hit rates obtained by NMR [52]. Other methods that also combine exhaustive sampling of the ligand-

receptor configurational space with severe approximations on the interaction energy predictions have proven useful to identify and characterize the most druggable binding sites of a target protein with a reasonable computational cost [53,54]. Obtaining quantitative predictions, however, requires more rigorous approaches that take into account often neglected terms such as solvation or entropy. This is achieved by the druggability index developed by Seco et al., which predicts the maximal binding affinity that a drug-like compound could achieve for a binding site from molecular simulations based on first principles [55]. Initially, the method reproduces a solvent-mapping experiment, in which the protein is exposed to a certain concentration of an organic solvent. Both NMR and crystallographic experiments have demonstrated that organic solvents tend to localize on binding sites[40,42,43], which is a natural consequence of the tendency of binding hot-spots to become desolvated (see above). Molecular dynamics simulations using 20% isopropyl alcohol (IPA) as solvent reproduce this behaviour, correctly identifying the experimentally determined IPA binding sites. Knowing that the method provides a correct sampling of the protein-ligand space, the collection of configurations generated by molecular dynamics can be subjected to a statistical treatment leading to binding site identification and druggability predictions. The process is illustrated in Fig. 2 and summarized here:

1. A grid encompassing the whole of the simulation box is generated and the number of times that a solvent atom type (IPA-OH, IPA-CH_3, Water-O) falls within each grid element is counted. Comparing the observed population (N_i) with the expected value (N_o), the associated free energy can be obtained using Eq. 1, where k_B is the Boltzmann constant and T the temperature at which the simulation was run.

   $$\Delta G_i = -k_B T \ln(Ni/No) \qquad (1)$$

2. The points with the best interaction free energies are identified, taking care that all points are separated by –at least– the distance of a covalent bond.
3. Points corresponding to IPA atom types (OH and CH_3) are considered transferable to aliphatic and polar neutral features of drug-like compounds, respectively. They are clustered together to form binding sites of maximal binding efficiencies.

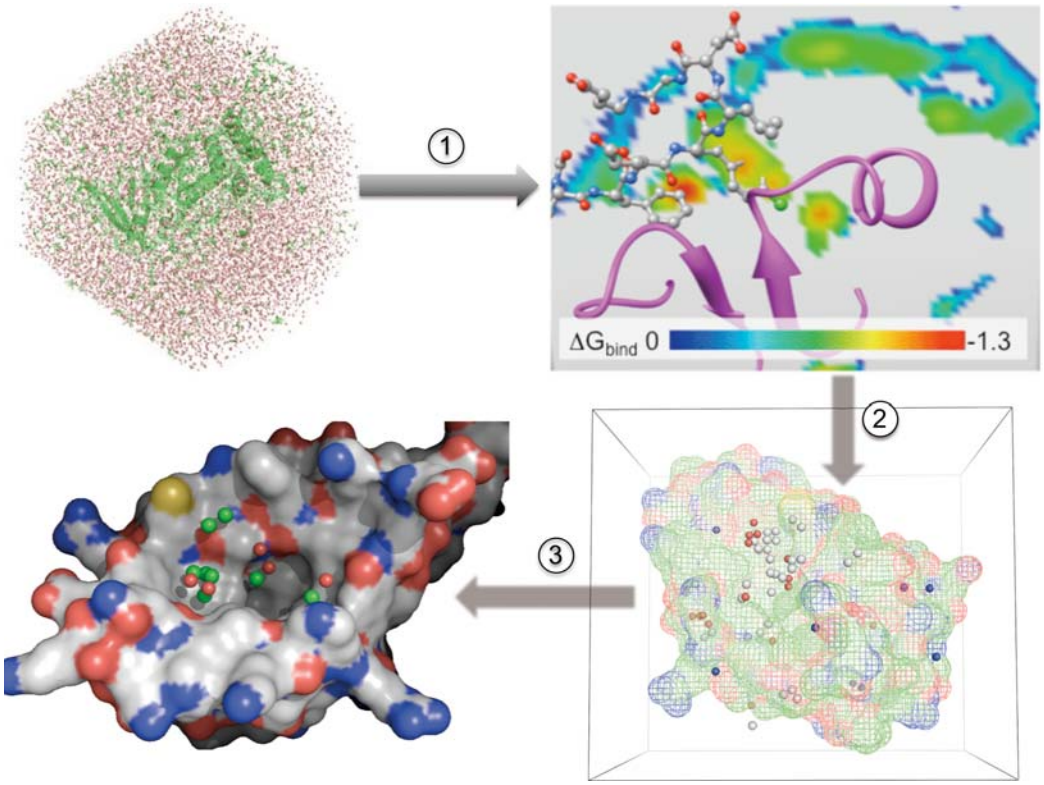


Figure 2. Detection of binding sites and estimation of the maximal binding free energy that a drug-like ligand can achieve, following the procedure by Seco et al. [55].

7 A TEST CASE: PTP1B

The protein phosphatase 1B (PTP1B) is a target for the treatment of type II diabetes and obesity that has proven extremely challenging. Many inhibitors acting on the phosphotyrosine binding site have been described [56], but potency is heavily dependent on the presence of a negative charge, which greatly damages its pharmacokinetics properties [57]. In consequence, its druggability classification is debatable: it has been considered druggable based on fragment screening hit rates [16] and success in hit identification motivated a sustained effort by many groups [58], but turning inhibitors into drugs has not been possible and an "undruggable" classification seems more appropriate. Empirical methods reproduce the prediction for which they have been trained, so it is considered druggable by Hajduk et al. [16] but undruggable by the other published methods [17,18,59]. A prediction based on first principle methods reveals that there is not a single hot spot for lipophilic or neutral polar features around the

phosphotyrosine binding site, which not only classifies it as undruggable, but also explains the total dependency of the charge to achieve potency [55].

Although the target is objectively difficult, development of an oral drug can never be ruled out. In fact, there have been two interesting developments that illustrate the importance of protein flexibility –one of the major challenges in drug design [60]– and the need to consider additional mechanisms of action. Two distinct conformations had been described for the so-called WPD loop (residues 179-184), which lines the catalytic site of PTP1B. In the apo form, this loop adopts an open conformation, whereas substrate binding induces a closing of the loop, thus reducing the size of the cavity that now fits tightly around the phosphotyrosine [61] (Fig. 3). Interestingly, this conformational change is coupled to a larger amplitude transition in the α7-helix (residues 287-295), located some 20Å away. In the WPD-closed conformation this helix is ordered and in contact with the α3-helix, but in the WPD-open form it is disordered and separated from the rest of the protein. Researchers at Sunesis discovered non-ionic inhibitors that

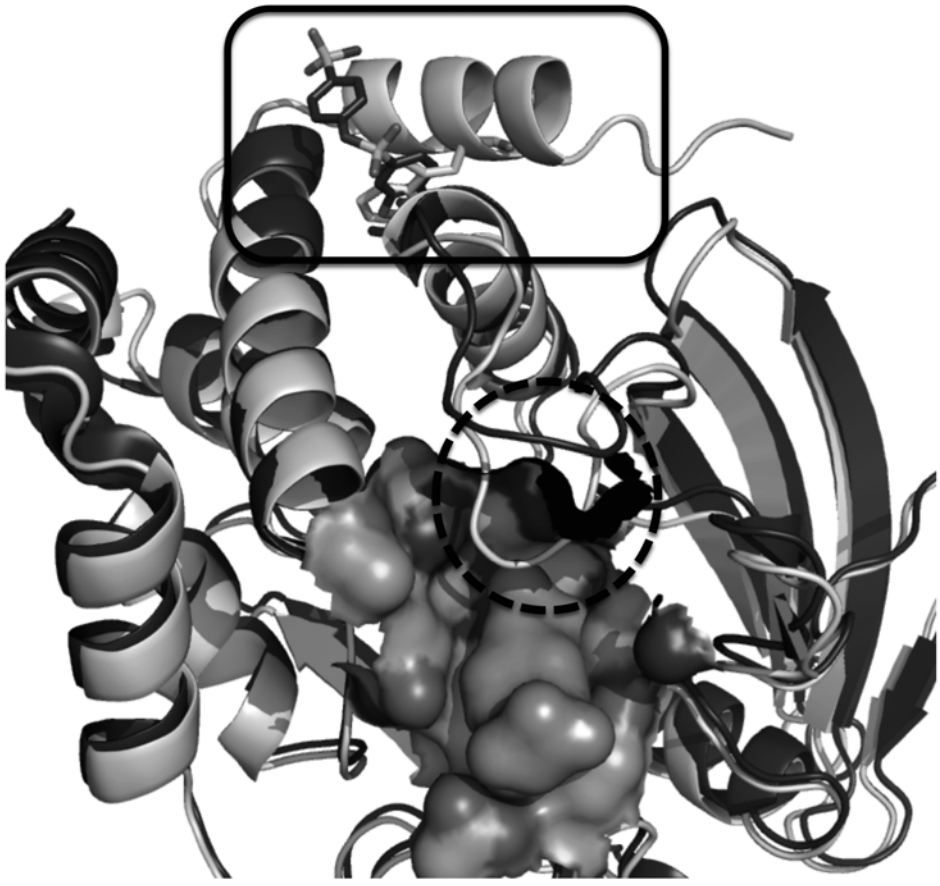


Figure 3. Superposition of active (pale grey) and inactive (dark grey) conformations of PTP-1B. In the active form, the catalytic site WPD loop (dashed circle) is closed and the α7-helix (box) is

packet against the rest of the protein. In the inactive form, the loop is open, leaving a large and shallow binding site. Allosteric inhibitors bind to a pocket that opens upon movement of the α7-helix, thereby overstabilizing the inactive conformation.

bind to a hydrophobic pocket that appears upon displacement of W291 (part of the α7-helix). Occupation of this pocket stabilizes the inactive WPD-open conformation and results in allosteric inhibition. In addition to providing a completely different chemotype with good cell permeation, the allosteric binding site is poorly conserved amongst phosphatases, making these compounds highly selective for PTP1B [62]. The druggability of this site is difficult to assess at present because, although the ligands are drug-like, they are weak binders (low μM). The empirical method based in fpocket identifies this binding site, but assigns a borderline druggability value [17] while the physics-based method predicts a maximal K_d of 500nM [55].

Very recently, the WPD-open conformation (inactive form) has been exploited to identify non-competitive inhibitors. Although they bind to the phosphotyrosine binding site, rather than competing with the substrate, they simply stabilize the inactive conformation, reducing the concentration of the catalytically competent enzyme, a mode of action known as conformational trapping [63]. Unlike most direct inhibitors, these molecules do not bear a negative charge, can cross membranes and achieve cellular activity [64].

8 OUTLOOK AND CONCLUDING REMARKS

Formal investigation of the causes of druggability has only started in recent years. Sitting at the interface of pharmacokinetics, molecular recognition and biomolecular structure, this incipient knowledge area builds on previous methods and understanding about drug-likeness, binding site identification and structure-based drug design, amongst others. Driven by a real necessity from the pharmaceutical industry, significant progress has been achieved. Of particular note is the existence of a small but diverse set of druggability prediction methods and the creation of a catalogue of systems with various degrees of druggability against which new methods can be trained and tested. Future challenges include explicit consideration of protein flexibility and achieving more quantitative and informative predictions.

Druggability prediction methods are expected to have two seemingly opposed consequences: on the one hand they will help concentrate on those targets offering better prospects, but they will also raise awareness about less obvious binding sites that may be used to exert a biological effect through non-standard mechanisms such as protein-protein inhibition [65], protein-protein

stabilization [66], target chaperoning [67], conformational trapping [63] and allosterism in general [26].

9 References

[1] C. Smith, Drug target validation: Hitting the target, *Nature,* **422,** 341, 343, 345 passim (2003).

[2] R. Macarron, Critical review of the role of HTS in drug discovery, *Drug Discov.Today,* **11,** 277-279 (2006).

[3] A.L. Hopkins and C.R. Groom, The druggable genome, *Nat.Rev.Drug Discov.,* **1,** 727-730 (2002).

[4] C.A. Lipinski, F. Lombardo, B.W. Dominy and P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv.Drug Deliv.Rev.,* **46,** 3-26 (2001).

[5] M.C. Wenlock, R.P. Austin, P. Barton, A.M. Davis and P.D. Leeson, A comparison of physiochemical property profiles of development and marketed oral drugs, *J.Med.Chem.,* **46,** 1250-1256 (2003).

[6] M. Vieth, M.G. Siegel, R.E. Higgs, I.A. Watson, D.H. Robertson, K.A. Savin, G.L. Durst and P.A. Hipskind, Characteristic physical properties and structural fragments of marketed oral drugs, *J.Med.Chem.,* **47,** 224-232 (2004).

[7] J.R. Proudfoot, The evolution of synthetic oral drug properties, *Bioorg.Med.Chem.Lett.,* **15,** 1087-1090 (2005).

[8] E. Perola, An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs, *J.Med.Chem.,* **53,** 2986-2997 (2010).

[9] C. Bissantz, B. Kuhn and M. Stahl, A medicinal chemist's guide to molecular interactions, *J.Med.Chem.,* **53,** 5061-5084 (2010).

[10] T. Clackson and J.A. Wells, A hot spot of binding energy in a hormone-receptor interface, *Science,* **267,** 383-386 (1995).

[11] W.L. DeLano, Unraveling hot spots in binding interfaces: Progress and challenges, *Curr.Opin.Struct.Biol.,* **12,** 14-20 (2002).

[12] A.A. Bogan and K.S. Thorn, Anatomy of hot spots in protein interfaces, *J.Mol.Biol.,* **280,** 1-9 (1998).

[13] F. Rodier, R.P. Bahadur, P. Chakrabarti and J. Janin, Hydration of protein-protein interfaces, *Proteins,* **60,** 36-45 (2005).

[14] I. Halperin, H. Wolfson and R. Nussinov, Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. implications for docking, *Structure,* **12,** 1027-1038 (2004).

[15] C. Sotriffer and G. Klebe, Identification and mapping of small-molecule binding sites in proteins: Computational tools for structure-based drug design, *Farmaco,* **57,** 243-251 (2002).

[16] P.J. Hajduk, J.R. Huth and S.W. Fesik, Druggability indices for protein targets derived from NMR-based screening data, *J.Med.Chem.,* **48,** 2518-2525 (2005).

[17] P. Schmidtke and X. Barril, Understanding and predicting druggability. A high-throughput method for detection of drug binding sites, *J.Med.Chem.,* **53,** 5858-5867 (2010).

[18] A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A.C. Salzberg and E.S. Huang, Structure-based maximal affinity model predicts small-molecule druggability, *Nat.Biotechnol.,* **25,** 71-75 (2007).

[19] A.R. Fersht, The hydrogen bond in molecular recognition, *Trends Biochem.Sci.,* **12,** 301-304 (1987).

[20] P.A. Karplus, Hydrophobicity regained, *Protein Sci.,* **6,** 1302-1307 (1997).

[21] J. Gao, D.A. Bosco, E.T. Powers and J.W. Kelly, Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins, *Nat.Struct.Mol.Biol.,* **16,** 684-690 (2009).

[22] C. Tang, J. Iwahara and G.M. Clore, Visualization of transient encounter complexes in protein-protein association, *Nature,* **444,** 383-386 (2006).

[23] A.P. Russ and S. Lampel, The druggable genome: An update, *Drug Discov.Today,* **10,** 1607-1610 (2005).

[24] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy and A. Bateman, The pfam protein families database, *Nucleic Acids Res.,* **38,** D211-22 (2010).

[25] L.Y. Han, C.J. Zheng, B. Xie, J. Jia, X.H. Ma, F. Zhu, H.H. Lin, X. Chen and Y.Z. Chen, Support vector machines approach for predicting druggable proteins: Recent progress in its exploration and investigation of its usefulness, *Drug Discov.Today,* **12,** 304-313 (2007).

[26] J.E. Lindsley and J. Rutter, Whence cometh the allosterome? *Proc.Natl.Acad.Sci.U.S.A.,* **103,** 10533-10535 (2006).

[27] D. Gonzalez-Ruiz and H. Gohlke, Targeting protein-protein interactions with small molecules: Challenges and perspectives for computational binding epitope detection and ligand finding, *Curr.Med.Chem.,* **13,** 2607-2625 (2006).

[28] A. Whitty and G. Kumaravel, Between a rock and a hard place? *Nat.Chem.Biol.,* **2,** 112-118 (2006).

[29] K. Venkatesan, J.F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.I. Goh, M.A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J.M. Sahalie, S. Cevik, C. Simon, A.S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R.R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M.E. Cusick, F.P. Roth, D.E. Hill, J. Tavernier, E.E. Wanker, A.L. Barabasi and M. Vidal, An empirical framework for binary interactome mapping, *Nat.Methods,* **6,** 83-90 (2009).

[30] A. Gupta, A.K. Gupta and K. Seshadri, Structural models in the assessment of protein druggability based on HTS data, *J.Comput.Aided Mol.Des.,*(2009).

[31] P. Ertl, Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups, *J.Chem.Inf.Comput.Sci.,* **43,** 374-380 (2003).

[32] S.B. Shuker, P.J. Hajduk, R.P. Meadows and S.W. Fesik, Discovering high-affinity ligands for proteins: SAR by NMR, *Science,* **274,** 1531-1534 (1996).

[33] M.L. Verdonk, V. Berdini, M.J. Hartshorn, W.T. Mooij, C.W. Murray, R.D. Taylor and P. Watson, Virtual screening using protein-ligand docking: Avoiding artificial enrichment, *J.Chem.Inf.Comput.Sci.,* **44,** 793-806 (2004).

[34] D.A. Erlanson, J.A. Wells and A.C. Braisted, Tethering: Fragment-based drug discovery, *Annu.Rev.Biophys.Biomol.Struct.,* **33,** 199-223 (2004).

[35] C. Dalvit, NMR methods in fragment screening: Theory and a comparison with other biophysical techniques, *Drug Discov.Today,* **14,** 1051-1057 (2009).

[36] S. Perspicace, D. Banner, J. Benz, F. Muller, D. Schlatter and W. Huber, Fragment-based screening using surface plasmon resonance technology, *J.Biomol.Screen.,* **14,** 337-349 (2009).

[37] M.M. Hann, A.R. Leach and G. Harper, Molecular complexity and its impact on the probability of finding leads for drug discovery, *J.Chem.Inf.Comput.Sci.,* **41,** 856-864 (2001).

[38] T. Fink, H. Bruggesser and J.L. Reymond, Virtual exploration of the small-molecule chemical universe below 160 daltons, *Angew.Chem.Int.Ed Engl.,* **44,** 1504-1508 (2005).

[39] I.J. Chen and R.E. Hubbard, Lessons for fragment library design: Analysis of output from multiple screening campaigns, *J.Comput.Aided Mol.Des.,*(2009).

[40] E. Liepinsh and G. Otting, Organic solvents identify specific ligand binding sites on protein surfaces, *Nat.Biotechnol.,* **15,** 264-268 (1997).

[41] C. Mattos and D. Ringe, Locating and characterizing binding sites on proteins, *Nat.Biotechnol.,* **14,** 595-599 (1996).

[42] A.C. English, C.R. Groom and R.E. Hubbard, Experimental and computational mapping of the binding surface of a crystalline protein, *Protein Eng.,* **14,** 47-59 (2001).

[43] C. Mattos, C.R. Bellamacina, E. Peisach, A. Pereira, D. Vitkup, G.A. Petsko and D. Ringe, Multiple solvent crystal structures: Probing binding sites, plasticity and hydration, *J.Mol.Biol.,* **357,** 1471-1482 (2006).

[44] S. Henrich, O.M. Salo-Ahen, B. Huang, F.F. Rippmann, G. Cruciani and R.C. Wade, Computational approaches to identifying and characterizing protein binding sites for ligand design, *J.Mol.Recognit.,* **23,** 209-219 (2010).

[45] S.J. Campbell, N.D. Gold, R.M. Jackson and D.R. Westhead, Ligand binding: Functional site location, similarity and docking, *Curr.Opin.Struct.Biol.,* **13,** 389-395 (2003).

[46] V. Le Guilloux, P. Schmidtke and P. Tuffery, Fpocket: An open source platform for ligand pocket detection, *BMC Bioinformatics,* **10,** 168 (2009).

[47] A. Tripathi and G.E. Kellogg, A novel and efficient tool for locating and characterizing protein cavities and binding sites, *Proteins,* **78,** 825-842 (2010).

[48] A. Volkamer, A. Griewel, T. Grombacher and M. Rarey, Analyzing the topology of active sites: On the prediction of pockets and subpockets, *J.Chem.Inf.Model.,* **50,** 2041-2052 (2010).

[49] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, The protein data bank, *Nucleic Acids Res.,* **28,** 235-242 (2000).

[50] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, DrugBank: A knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Res.,* **36,** D901-6 (2008).

[51] Chipot, C. & A. Pohorille. Free Energy Calculations: Theory and Applications in Chemistry and Biology. A. W. J. Castleman, J. P. Toennies, K. Yamanouchi and W. Zinth(eds). In "Springer Series in Chemical Physics". Berlin:Springer (2007).

[52] N. Huang and M.P. Jacobson, Binding-site assessment by virtual fragment screening, *PLoS One,* **5,** e10109 (2010).

[53] R. Brenke, D. Kozakov, G.Y. Chuang, D. Beglov, D. Hall, M.R. Landon, C. Mattos and S. Vajda, Fragment-based identification of druggable 'hot spots' of proteins using fourier domain correlation techniques, *Bioinformatics,* **25,** 621-627 (2009).

[54] M. Clark, F. Guarnieri, I. Shkurko and J. Wiseman, Grand canonical monte carlo simulation of ligand-protein binding, *J.Chem.Inf.Model.,* **46,** 231-242 (2006).

[55] J. Seco, F.J. Luque and X. Barril, Binding site detection and druggability index from first principle, *J. Med. Chem.,* **52,** 2363-2371 (2009).

[56] T.N. Doman, S.L. McGovern, B.J. Witherbee, T.P. Kasten, R. Kurumbail, W.C. Stallings, D.T. Connolly and B.K. Shoichet, Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B, *J.Med.Chem.,* **45,** 2213-2221 (2002).

[57] C. Abad-Zapatero, Ligand efficiency indices for effective drug discovery, *Expert Opin.Drug Discov.,* **2,** 469-488 (2007).

[58] B. Kasibhatla, J. Wos and K.G. Peters, Targeting protein tyrosine phosphatase to enhance insulin action for the potential treatment of diabetes, *Curr.Opin.Investig Drugs,* **8,** 805-813 (2007).

[59] T.A. Halgren, Identifying and characterizing binding sites and assessing druggability, *J.Chem.Inf.Model.,* **49,** 377-389 (2009).

[60] P. Cozzini, G.E. Kellogg, F. Spyrakis, D.J. Abraham, G. Costantino, A. Emerson, F. Fanelli, H. Gohlke, L.A. Kuhn, G.M. Morris, M. Orozco, T.A. Pertinhez, M. Rizzi and C.A. Sotriffer, Target flexibility: An emerging consideration in drug discovery and design, *J.Med.Chem.,* **51,** 6237-6255 (2008).

[61] Z. Jia, D. Barford, A.J. Flint and N.K. Tonks, Structural basis for phosphotyrosine peptide recognition by protein tyrosine phosphatase 1B, *Science,* **268,** 1754-1758 (1995).

[62] C. Wiesmann, K.J. Barr, J. Kung, J. Zhu, D.A. Erlanson, W. Shen, B.J. Fahr, M. Zhong, L. Taylor, M. Randal, R.S. McDowell and S.K. Hansen, Allosteric inhibition of protein tyrosine phosphatase 1B, *Nat.Struct.Mol.Biol.,* **11,** 730-737 (2004).

[63] G.M. Lee and C.S. Craik, Trapping moving targets with small molecules, *Science,* **324,** 213-215 (2009).

[64] S. Liu, L.F. Zeng, L. Wu, X. Yu, T. Xue, A.M. Gunawan, Y.Q. Long and Z.Y. Zhang, Targeting inactive enzyme conformation: Aryl diketoacid derivatives as a new class of PTP1B inhibitors, *J.Am.Chem.Soc.,* **130,** 17075-17084 (2008).

[65] M.R. Arkin and J.A. Wells, Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream, *Nat.Rev.Drug Discov.,* **3,** 301-317 (2004).

[66] Y. Pommier and J. Cherfils, Interfacial inhibition of macromolecular interactions: Nature's paradigm for drug discovery, *Trends Pharmacol.Sci.,* **26,** 138-145 (2005).

[67] P. Leandro and C.M. Gomes, Protein misfolding in conformational disorders: Rescue of folding defects and chemical chaperoning, *Mini Rev.Med.Chem.,* **8,** 901-911 (2008).

Large-Scale Comparison of Four Binding Site Detection Algorithms

Peter Schmidtke, Catherine Souaille, Frédéric Estienne, Nicolas Baurin, and Romano T. Kroemer

Journal of Chemical Information and Modelling, 2010, 50

Large-Scale Comparison of Four Binding Site Detection Algorithms

Peter Schmidtke,[†,‡] Catherine Souaille,[†] Frédéric Estienne,[†] Nicolas Baurin,[†] and Romano T. Kroemer*[,†]

Sanofi-Aventis VA Research Centre, Structure Design & Informatics, 13 quai Jules Guesde, BP14, 94403 Vitry-sur-Seine, France and Departament de Fisicoquimica and Institut de Biomedicina (IBUB), Facultat de Farmacia, Universitat de Barcelona, 08028, Barcelona, Spain

Received January 18, 2010

A large-scale evaluation and comparison of four cavity detection algorithms was carried out. The algorithms SiteFinder, fpocket, PocketFinder, and SiteMap were evaluated on a protein test set containing 5416 protein−ligand complexes and 9900 apo forms, corresponding to a subset of the set used earlier for benchmarking the PocketFinder algorithm. For the holo structures, all four algorithms correctly identified a similar amount of pockets (around 95%). SiteFinder, using optimized parameters, SiteMap, and fpocket showed similar pocket ranking performance, which was defined by ranking the correct binding site on rank 1 of the predictions or within the first 5 ranks of the predictions. On the apo structures, PocketFinder especially and also SiteFinder (optimized parameters) performed best, identifying 96% and 84% of all binding sites, respectively. The fpocket program predicts binding sites most accurately among the algorithms evaluated here. SiteFinder needed an average calculation time of 1.6 s compared with 2 min for SiteMap and around 2 s for fpocket.

INTRODUCTION

The number of known protein three-dimensional (3D) structures in the public and private domains is constantly on the rise. For drug discovery purposes these structures are of great interest, as they can be exploited in the search for small molecules that bind to them and modulate their function. Of particular importance are the cavities at the protein surface, as they provide the best environment for anchoring small molecules.

In many cases cavities can be identified by the presence of natural substrates, a cofactor or a ligand. However some proteins are crystallized without any partner. Detecting cavities at the surface of these proteins can help in finding the natural substrate binding site, identifying binding pockets or allosteric sites to start the design of small-molecule ligands of therapeutic effect. Moreover, given a cavity in one protein, the detection of similar cavities in other proteins may provide hints to anticipate issues, such as selectivity or toxicity.

A first step toward exploitation and comparison of pockets as well as cavity-based annotation of proteins is therefore the comprehensive scanning of protein 3D structures with the aim of detecting all cavities of interest. These cavities can subsequently be analyzed and compared with novel programs, such as SuMo[1] or FLAP.[2] Given that these programs take surface shape and property into account, they are well suited for the task of cavity/pocket comparison, as opposed to programs that analyze protein backbone and topology, such as SARF2,[3] VAST,[4,5] DALI,[6,7] and FATCAT.[8,9]

Over the past years a number of different approaches have been developed to correctly predict binding sites on the protein surface. One can distinguish two different types of cavity finding algorithms: (i) evolutionary- and (ii) structure-based algorithms. The second category can be subdivided in geometry- and energy-based algorithms.

The most popular example of geometry-based algorithms in the public domain is putative active sites with spheres (PASS).[10] Other well-known geometry-based algorithms are SURFNET[11] and LIGSITE[12,13] (improved version of POCKET).[14] APROPOS[15] and CAST[16] are based on alpha shape analysis.[17,18] The recently published fpocket[19] uses similar properties derived as alpha spheres, already employed by the SiteFinder[20] algorithm.

Energy-based algorithms like PocketFinder,[21] the method introduced by Bliznyuk and Gready,[22] the computational mapping from the Vajda group,[23,24] the multiscale approach from Glick,[25] and the method developed by Ruppert,[26] or SuperStar[27] simulate the interactions of a solvent molecule on the protein surface in order to detect local surface properties of a cavity. Some of these methods still use a geometry-based step in order to measure the extent of the cavity, by tracing rays from grid points in the cavity. Nevertheless approaches like the multiple solvent mapping developed by the Vajda group are fully based on interaction energy calculations.

The accuracy of most of the cavity finding algorithms has not been evaluated on large data sets. Only the PocketFinder algorithm[21] published in 2005 provided a large-scale evaluation using a data set of 17.626 proteins from the Protein Data Bank (PDB). This evaluation included an assessment of the algorithms' capacity to recognize binding sites on apo forms as compared to the corresponding ligated proteins.

In the present study, a large scale evaluation of four cavity finding algorithms has been carried out. Two of them are implemented in two major molecular modeling packages:

* Corresponding author. E-mail: romano.kroemer@sanofi-aventis.com.
[†] Sanofi-Aventis VA Research Centre, In Silico Sciences/Drug Design, 13 quai Jules Guesde, BP14, 94403 Vitry-sur-Seine, France.
[‡] Departament de Fisicoquimica and Institut de Biomedicina (IBUB), Facultat de Farmacia, Universitat de Barcelona, 08028, Barcelona, Spain.

10.1021/ci1000289
Published on Web 09/09/2010

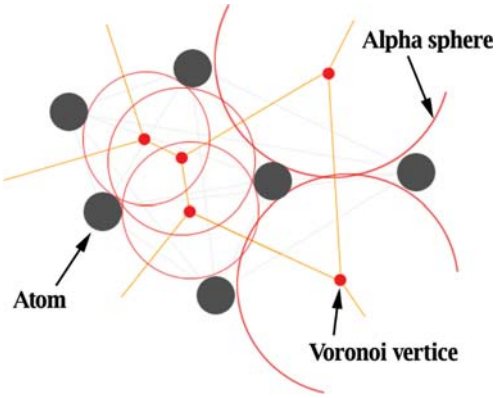


Figure 1. Example of an alpha sphere. The alpha sphere is displayed in red, and the three contacted atoms (2D) are in gray.

SiteFinder, which is implemented in the Molecular Operating Environment (MOE) software provided by the Chemical Computing Group (CCG),[28] and SiteMap,[29,30] provided by Schrödinger[31] and accessible through the Maestro graphical user interface or through the command line. To our knowledge, no evaluation of SiteFinder has been published up to now and a data set of 297 proteins[29,30] has been used for the development and optimization of SiteMap. Furthermore, PocketFinder and fpocket are compared to the previous two. PocketFinder, falling also in the category of energy-based algorithms, was chosen because it had previously been evaluated on the data set used in this study. Fpocket, a geometry-based algorithm, similar to SiteFinder, was included as a sole open source alternative to the previously cited methods.

The main criteria in the evaluation of all algorithms was success rate and accuracy of binding site identification as well as computational performance. Ease-of-use, such as scripting facilities and accessibility of results, were also considered. In the following, the methods evaluated are being described in more detail.

SiteFinder. SiteFinder falls into the category of geometric methods, since no energy models are used. Relative positions and accessibility of the receptor atoms are considered along with an approximate classification of chemical type. The method is based on the identification of regions of tight atom packing on the protein, filtering of exposed regions, hydrophobic/hydrophilic classifications, and the use of a definition of hydrophilicity that is invariant to protonation state.

No grid-based method is used for SiteFinder, this way the method is invariant to rotation of atomic coordinates, and less memory is required for the calculation.

The SiteFinder methodology is based upon alpha shapes which are a generalization of convex hulls developed by Edelsbrunner.[17] A collection of 3D points is triangulated using a modified Delaunay triangulation. For each resulting simplex (collection of four points), there is an associated sphere called alpha sphere (Figure 1). These spheres have different radii.

The collection of alpha spheres is pruned by eliminating those that correspond to inaccessible regions of the receptor as well as those that are too exposed to solvent. In addition, only small alpha spheres are retained since these correspond to locations of tight atom packing in the receptor. Each alpha sphere is classified as either "hydrophobic" or "hydrophilic", depending on whether the sphere is in a good hydrogen-bonding spot on the receptor. Hydrophilic spheres that are far from hydrophobic spheres are eliminated. All alpha spheres are clustered using a single linkage clustering algorithm. A key feature is that each cavity consists of one or more alpha spheres and at least one hydrophobic alpha sphere. Resulting cavities are ranked using the number of contacts with hydrophobic atoms of the receptor.

fpocket. The fpocket algorithm is based on very similar principles as SiteFinder. Using Voronoi tessellation through the free computational geometry library Qhull, fpocket filters Voronoi vertices and their corresponding alpha spheres according to alpha sphere minimum and maximum radii. The next three clustering steps are performed to aggregate nearby alpha spheres to form a pocket (set of alpha spheres). Each cluster of alpha spheres forms a putative pocket. Each pocket is scored based on a knowledge-based SiteScore, and the final pocket list is ranked using this score.

Compared to SiteFinder, fpocket is currently a command line driven open source cavity detection algorithm. Next to basic pocket prediction, fpocket integrates several tools for easy extraction of pocket descriptors and for testing scoring function. A druggability prediction score has recently been integrated as well.[32]

SiteMap. SiteMap is an energy-based cavity finding algorithm. It identifies probable binding sites through three main steps: (i) detection of cavities, (ii) characterization of detected cavities, and (iii) evaluation of characterized cavities.

In the first step, a 1 Å grid of site points is built around the entire protein; points overlapping the protein atoms are deleted. Then the algorithm filters out the site points located too far from the protein or displaying a low degree of "enclosure" within the receptor. The "enclosure" of site points is computed using rays traced in all directions. The number of rays cutting the protein surface at a certain distance is used to estimate the relative "enclosure" of the grid point. The points that fulfill these criteria are clustered into site point groups. Groups of site points are merged if the distance between them is below a predefined threshold and occurs in a solvent-exposed region. The default maximum distance between two grid points to be merged into the same group is 6.5 Å. The ratio of the distance between the centroids of the groups to their effective size (default value is 5) determines whether the groups are considered for merging.

During the second step—the mapping process—various properties of the cavity are calculated using the remaining site points. Hydrophobic and hydrophilic potentials are generated using van der Waals and electric field grids. SiteMap then partitions the accessible space in each site into hydrophobic, hydrophilic, and "neither/nor" regions. The hydrophilic map is further divided into hydrogen-bond donor and acceptor maps. The last step of the SiteMap cavity detection procedure is the site evaluation. Various scores are calculated by SiteMap. The main score (SiteScore) is based on a weighted sum of the following criteria:

- Number of site points: the number of grid points necessary to define the cavity.
- Exposure/enclosure: the property to measure how open the cavity is to the solvent.

- Hydrophobic/hydrophilic character and balance: the measure of the relative hydrophobic/hydrophilic character of the cavity. Balance is the ratio hydrophobic/hydrophilic score.
- Donor/acceptor character: the estimated hydrogen-bond donor intensity of a putative ligand in the cavity.

PocketFinder. The algorithm, published by An et al. in 2005, falls into the category of energy-based pocket detection algorithms. It makes use of a transformed Lennard-Jones potential calculated on the protein structure using an aliphatic carbon atom as a probe placed on a 1.0 Å spaced grid over the protein. Next, the grid is smoothed to emphasize regions with a consistently low Lennard-Jones potential over a given region in space. Out of these isolated regions, envelopes are created and further filtered for envelopes having a volume bigger than 100 $Å^3$. Finally, the resulting binding sites are ranked by volume.

MATERIALS AND METHODS

To carry out the comparison of the algorithms, a two-step procedure was applied. First, a preliminary study was performed on a small calibration data set with a view to eventually optimizing some of the parameters determining the cavity detection. Second, the algorithms with the assessed search parameters were used to check their ability to correctly identify binding sites on a large evaluation data set.

Preliminary Study. The calibration data set consists of 370 protein−ligand complexes obtained by X-ray diffraction with a resolution better than 2.5 Å. These structures are part of a MOE sample database (complex.mdb file shipped with MOE) with structures cleaned from water and crystallization additives, which could alter search results on the protein surfaces. No hydrogen atoms were considered for pocket detection with SiteFinder. Both cavity finding algorithms were applied in the absence of the ligand in the investigated binding site and tested for their performance on the following criteria:

- Percentage of found binding sites ranked as first (according the algorithm score).
- Percentage of found binding sites ranked within the first five positions.

The aim of the calibration step was to minimize: (i) the percentage of not found binding sites; (ii) the percentage of binding sites split up in multiple pockets, and (iii) the mean rank and standard deviation of the rank of a correctly detected binding site.

SiteFinder. Optimization was performed on three parameters of the SiteFinder algorithm (default values in brackets):

- Connect Dist (2.5 Å): connection distance between two alpha spheres, used to cluster alpha spheres into a common group (cavity).
- Minrad (2.0 Å): minimum threshold distance between the alpha spheres of a cluster and the centroid of this cluster.
- Da Dist (3.0 Å): maximum distance between hydrophilic alpha spheres and the nearest hydrophobic alpha sphere.

Connect dist and minrad, tuning the clustering procedure, were optimized with a combinatorial procedure. Da dist was thereafter optimized using the optimum values of the former parameters.

In addition, small cavities can be filtered out with the site minsize (3) parameter, referring to the number of alpha spheres in a site.

SiteMap. This algorithm makes use of 16 parameters that can be modified. The results obtained in the preliminary study with the default values of these parameters were satisfying. Moreover calculations on 370 structures using a single set of parameters took several days on a two-processor Linux workstation, thus it was not feasible performing an exhaustive parameter optimization within a reasonable amount of time. Therefore no parameter optimization was undertaken by us.

The MOE database previously compiled for the evaluation of SiteFinder was used to export receptor and ligand, each into separate files in PDB format. PDB receptor files were further converted into MAE file format (pdbconvert script provided by Schrödinger). Hydrogen atoms were then added to the receptor (applyhtreat script provided by Schrödinger), followed by an atom-typing step required to perform the energy calculation.

SiteMap does not handle structures with missing atoms. In order to overcome this issue, PrimeFill (Schrödinger) was used to build and refine the missing protein regions on the calibration data set. This process was however too long to be used on the evaluation data set. Thus, the evaluation step of SiteMap was performed only on structures prepared automatically using the prepwizard program provided by Schrödinger. Structures that could not be treated by SiteMap were excluded from the analysis for all algorithms.

No evaluation of the influence of different protonation states on the pocket prediction results is performed here, as one of the main objectives of this evaluation is to assess the suitability of SiteMap for high-throughput pocket prediction.

fpocket. No further calibration was carried out for fpocket.

PocketFinder. Also for this algorithm, no calibration was performed, as the results published by An et al.[21] were taken for this comparison.

Evaluation Study. The evaluation data set consisted initially of the 17 126 structures, with a resolution lower than 2.5 Å, generated to evaluate PocketFinder and to compute the so-called Pocketome.[21] When the present study was performed this data set was available on the Web site http://abagyan.ucsd.edu/index.html. It is the largest data set ever used to evaluate cavity finding algorithms. To our knowledge, so far, only PocketFinder has been evaluated with this data set. The set contains 5616 protein−ligand complexes and 11 510 apo structures on which the binding site location is known. At least one holo structure can be found among the 5616 complexes for each apo structure, thus enabling to assess whether the experimentally observed cavities in the complex structures can also be found in the apo structures.

An et al. used data deposited in the PDB from October 30, 2003. Between this date and today, major changes have been made for a multitude of the PDB structures in this data set. For some structures, PDB accession codes simply changed. For others, the known ligand molecule identifier changed. Also, a few PDB structures have been deleted from the PDB since then. For all the previously cited reasons, only 5416 structures out of the initial 5616 holo structures were retained for this study. New chain assignments and the previously cited reasons reduced the apo data set to 9900 structures. This data set reduction is also due to the fact that for example SiteMap could not run successfully (in an

automated manner) on some structures. Thus, to guarantee integrity of the whole data set for the different methods, the data set size was reduced by the structures that SiteMap cannot analyze. The final data set is provided as additional text files (c.f. Supporting Information).

First, a test on the 5416 holo proteins was performed in order to evaluate the ability of the four algorithms to identify the ligand binding site. Next, a second test was performed on the 9900 apo structures.

The criteria to evaluate the performance of both algorithms are:

- Percentage of correctly identified binding sites.
- Accuracy of the prediction.
- Correlation between results obtained on the holo and apo forms of the protein structures.
- Calculation time.
- Ease of use.

In the present analysis, structurally important cofactors, like hemes, biopterins, or chlorophyls, were considered as part of the receptor in order to build a functional unit. Thus these cofactors were included during binding site search. As some of the small molecules in the initial data set by An et al. correspond to this category of molecules, these structures were taken out of the data set. The final data set used on all algorithms is provided as csv files in the Supporting Information.

Definition of a Correctly Identified Ligand Binding Site. The definition of when a binding site is correctly identified is crucial in evaluations, such as the ones carried out here. In order to obtain results comparable to An et al., the same evaluation criterion as in the PocketFinder evaluation was used. This criterion, called relative overlap (RO) is defined as follows: $RO = (A_L \cap A_E)/A_L)$, where A_L is the solvent-accessible area of the receptor atoms within 3.5 Å from a bound ligand, and A_E is the solvent-accessible area of the receptor atoms within 3.5 Å from the predicted pocket envelope. A total mis-prediction would have $RO = 0$, whereas a perfect prediction would have $RO = 1.0$.

Whether the algorithm has detected correctly, a ligand binding site is determined by the overlap between the surface of the cavity atoms and the surface of the actual binding site atoms. If the RO is at least about 0.5, then the cavity will be assessed as "correctly identified".

In apo structures, the ligand binding site is defined by analogy with the holo structure after protein superposition.

Assessing the Accuracy of the Pocket Prediction. An et al. used the RO criterion to assess the accuracy of the pocket prediction. However, the accuracy of a prediction can be defined in various ways. The inherent disadvantage of using a criterion like the RO is that the bigger a predicted pocket gets, the more chance one has to reach a RO close to one, covering completely the known ligand binding site. If the purpose of a pocket prediction algorithm is however to propose reasonably sized binding pockets, then the RO alone is not enough to assess the accuracy of a prediction. Thus, a second criterion, called mutual overlap criterion (MOC) is introduced in this study. It is based on the same principle as the RO but is defined as: $MO = (A_L \cap A_E)/A_E)$. Similarly to RO, if the MO gets close to 1, then it is an indicator the predicted pocket is covering only the surface of the actual overlap between the ligand and predicted binding sites. Associating both RO and MO, one could get an estimate of the accuracy of pocket prediction.

Comparing Geometry-Based Methods with Energy-Based Methods. This evaluation intends to compare two energy-based pocket identification algorithms (SiteMap and PocketFinder) with two geometry-based algorithms (fpocket and SiteFinder). As the pockets are represented in very different ways such a comparison is not straightforward. Both PocketFinder and SiteMap use a grid to delimit the binding site. Thus a binding site can be represented as either an envelope or a set of grid points in the pocket. SiteFinder and fpocket produce more sparsely spaced alpha spheres. In order to assess if a binding site was correctly identified, the solvent accessible surface area of the pocket has to be calculated, and this can be done using a 3.5 Å distance from all grid points of the pocket or all alpha sphere centers. As grid points are more densely packed than alpha sphere centers, this would result in an underestimation of the correctly identified pocket surface for geometry-based methods.

In order to address this, a grid intended to be very similar to the SiteMap grid is packed into the alpha spheres of fpocket and SiteFinder. The general amber force field (GAFF) was used to assign van der Waals parameters to all atoms of the protein. Next, a 0.7 Å spaced grid is placed over the pocket. In these calculations, the radius and the well depth of the Lennard-Jones probe particles are taken to be 1.5 Å and 0.13 kcal/mol, respectively, as used by Halgren in SiteMap.[30] Finally, only those grid points within the alpha spheres were retained that are equal or further than the closest van der Waals equilibrium distance of the probe in the pocket. In all subsequent surface calculations, these retained grid points represent the pocket, on the contrary to the previously used alpha sphere centers (and volume).

PocketFinder uses a very similar representation of pockets and results obtained with SiteMap, and both geometry-based methods using the transformation presented in the previous paragraph are thus considered comparable.

Scripting and Statistical Analysis. Automation of cavity finding was scripted using the programming environment of each molecular modeling package. The Scientific Vector Language (SVL) was used in MOE (version 2006−2008). This scripting language is the proprietary MOE language and provides a flexible platform for users willing to develop their own methods.

In the Maestro molecular modeling suite, the Maestro Command Language is interfaced with Python. Versions 7.5 and 8.0 of Maestro were used in this study (see Results Section).

Analysis was performed using R statistical software version 2.2.1[33] and SpotFire DecisionSite 8.0.

All calculations were performed on biprocessor (2 × 3.6 Ghz) and 2 Gb RAM workstations, running under a RHEL WS release 3 distribution.

Results for pocket prediction for fpocket were taken from a precomputed pocket database (in-house). The transformation from alpha sphere-based pockets to grid-based pockets was performed using several Python based in-house libraries.

RESULTS

SiteFinder Preliminary Study. In order to evaluate SiteFinder with its maximum performance, all search pa-

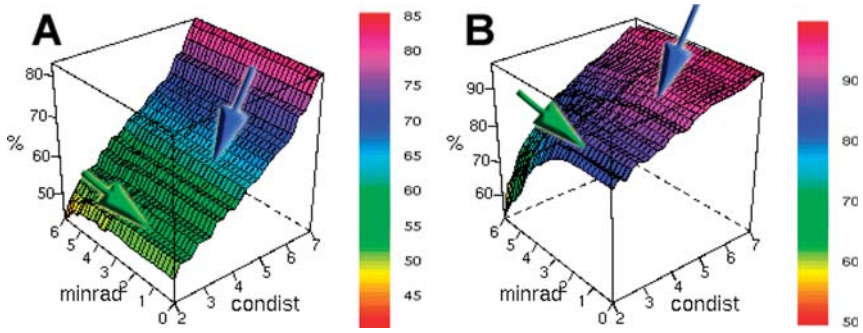


Figure 2. Optimization of SiteFinder parameters. (A) % of correctly identified binding sites ranked on the first rank in function of minrad and connect dist. (B) % of correctly identified binding sites found on the first five ranks in function of minrad and connect dist. Green and blue arrows represent default and optimized parameters, respectively.

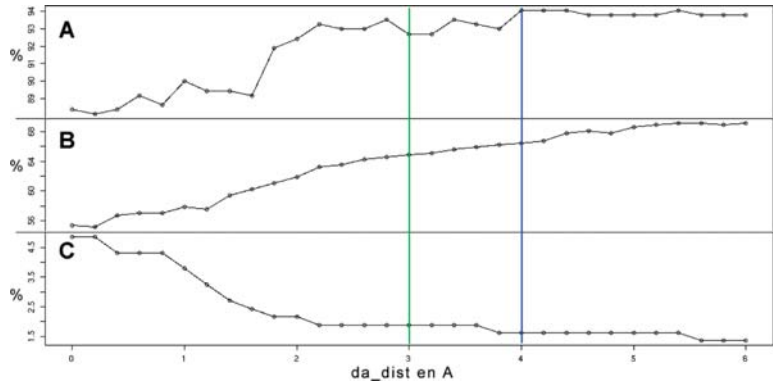


Figure 3. Optimization of SiteFinder parameters. (A) % of correctly identified binding sites ranked within the first five ranks as a function of da dist. (B) % of correctly identified binding sites ranked on the first rank as a function of da dist. (C) % of binding site not found by SiteFinder as a function of da dist.; Green and blue lines represent default and optimized parameters, respectively.

rameters were first optimized during the preliminary study. These calculations led to notable adjustments of minrad, connect dist and da dist values. Minrad was modified from 2.0 (default) to 1.8 Å(optimized), connect dist from 2.5 (default) to 4.6 Å(optimized), and da dist from 3.0 (default) to 4.0 Å(optimized).

In Figure 2A notable increase of correctly identified binding sites ranked on the first rank is observed with increasing connect dist value. Using default parameters (in brackets optimized parameters), SiteFinder ranks about 54 (65%) of identified binding sites on the first rank, 86 (92.7%) within the 5 first ranks. SiteFinder, with default parameters was not able to identify 3.2 (1.6%) of all binding sites during this optimization run on 370 protein structures. This optimization step was performed using a fixed value of 5 for site minsize parameter. Figures 2 and 3 show that cavity detection and ranking performance could be theoretically further enhanced by increasing the connect dist and the da dist values. However, care must be taken with further increasing these values. Considering, for instance, the connect dist parameter with its optimized value fixed at 4.6 Å, a further increase of this value would result in one single cavity at the protein surface. Therefore, connect dist and da dist were not increased further for the purposes of this study.

SiteFinder Full-Scale Evaluation. First, the evaluation step on SiteFinder was performed on the 5416 protein−ligand complexes full data set using default (in brackets optimized) parameters. Figure 4 illustrates that more than 75 (95%) of all binding sites are identified with a RO close to 1. At the threshold of a well-identified binding site, 95 (98%) of all binding sites are found. As shown in Figure 5, 70 (77%) of all found binding sites are ranked as first. Considering all binding sites found on ranks 1−5, the total would amount to 95 (98%) of all found binding sites.

The impressive increase in the percentage of found binding sites with an RO near 100% from SiteFinder with default parameters to the percentage found with optimized parameters shows that the RO alone is not a good enough criterion to evaluate the accuracy of a binding site prediction algorithm. Figure 6 depicts the MO of both parameter sets for SiteFinder, and one can observe a clear shift in accuracy from default to optimized parameters. This important shift toward lower MO values for optimized parameters clearly indicates that, although the RO is very high for most of the found binding sites, this comes at the cost of prediction of far too big binding sites.

Second, SiteFinder using default parameters (optimized parameters in brackets) was evaluated on 9900 apo structures. Here, more contrasted results were obtained. The algorithm was able to retrieve around 40 (65%) of all binding sites with a RO near 1, as shown in Figure 4. This corresponds to a drop of 35 (20%) compared to results obtained on holo structures. Also for ranking performance, a drop in the predictive power can be observed. SiteFinder ranks 42 (62%) of all found binding sites on rank 1 and 83 (98%) within the top five ranks.

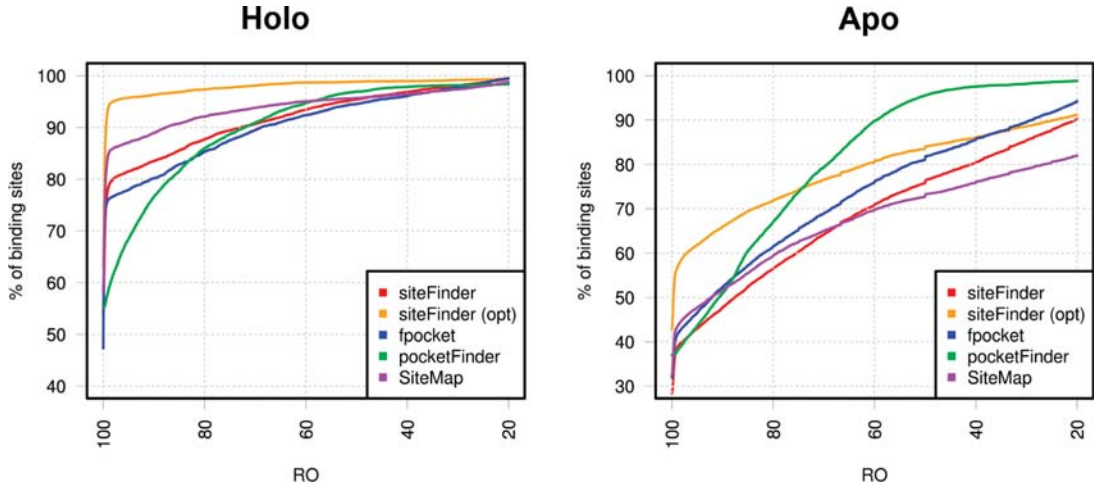


Figure 4. The prediction accuracy measured by the RO between the predicted binding patch A_E, defined as the solvent-accessible surface of the receptor atoms within 3.5 Å from the predicted envelope, and the observed binding patch A_L, defined as the solvent-accessible surface of the receptor atoms within 3.5 Å from the bound ligand. The results of 5416 binding sites from protein−ligand complexes and 9900 binding sites from uncomplexed structures were sorted separately by RO. SiteFinder using default parameters is red; SiteFinder with optimized parameters is orange; fpocket with default parameters is blue; PocketFinder (taken from An et al.)[21] is green; and SiteMap using default parameters is purple.

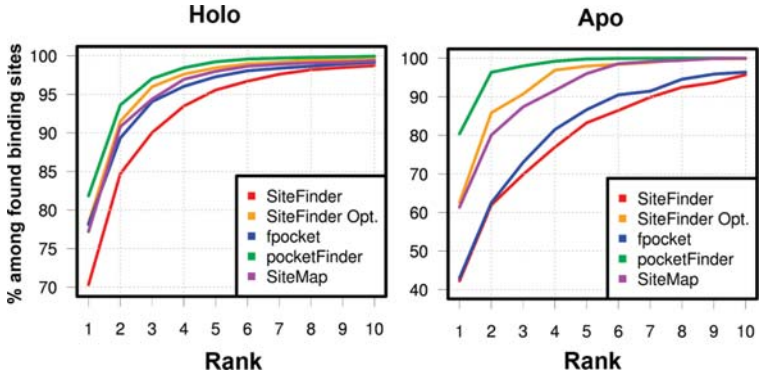


Figure 5. Cumulative percentage of binding sites among found binding sites (RO > 0.5) versus the ranking of those. For all methods, more than 95% of the identified binding sites are found among the first 5 ranks on holo structures.

The average calculation time per protein for SiteFinder was 1.6 s.

SiteMap Preliminary Study. During the preliminary study, the systematic binding site search was performed on 315 proteins out of 370. By default, SiteMap returns only the first five cavities. All binding sites from the 315 proteins were in this set of cavities. A total of 68% of actual binding sites were ranked as first, while 87% of them bind ligands with molecular weights (MW) larger than 250. Thus the SiteMap scoring function used to rank identified cavities performed well for our purpose, as all actual binding sites were retrieved in all cases. Also, the results indicated that no optimization of the search parameters of SiteMap was necessary.

The SiteMap process stopped during the atom-typing step for the remaining 55 proteins, with incomplete coordinates. Missing residues were modeled using Prime, a Schrödinger module that performs homology modeling and side chain and loop prediction. However this process was very time-consuming and did not succeed for all proteins, so it was not applied to the structures used in the full-scale study.

SiteMap Full-Scale Evaluation. As alluded to in the Materials and Methods Section, because SiteMap can handle only complete structures (no missing atoms or residues), structures that cannot be treated by SiteMap in an automated manner (the prepwizard program from Schrödinger was used to prepare the structures) were omitted from the data set. By default, the output of the SiteMap algorithm is limited to the five top-ranked cavities. To enable a relevant comparison with SiteFinder, the SiteMap output set was enlarged to 20 cavities. It should be noted that this modification had an influence on the cavity delimitation, resulting in some cases in the splitting of large cavities.

Figure 4 illustrates that around 85% of all binding sites were found with a relative overlap close to 1. At the threshold of a correctly identified binding site (RO > 0.5) SiteMap gave good predictions for 95% of all binding sites. Considering the ranking performance (Figure 5) of the score implemented in SiteMap, it managed to retrieve around 78% of found binding sites on the very first rank and around 97% within the top five ranks.

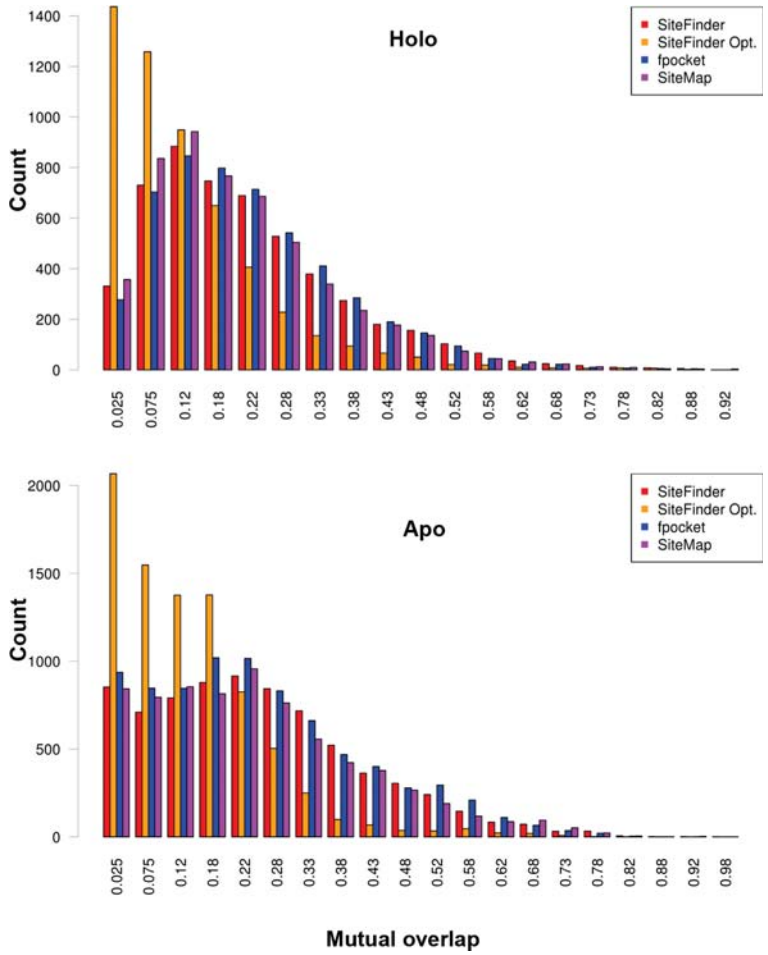


Figure 6. Introduction of mutual overlap, a second measure of prediction accuracy, entitled MO. MO is defined as the ratio between the overlapping ASA between predicted and known pockets and the total predicted pocket ASA. The higher the MO, the more accurately the pocket is predicted. SiteFinder using default parameters is red; SiteFinder using optimized parameters is orange; fpocket using default parameters is blue; and SiteMap using default parameters is purple.

As illustrated in Figures 4 and 5, SiteMap allowed to predict about 73% of all 9900 apo binding sites correctly, while 61% of these were ranked on rank 1 and 96% within the top five ranks.

The preliminary calculations were performed with Maestro 7.5 suite, and the average calculation time was about 13 min per structure, including system preparation. Variations in calculation time were rather large, from 2 min to several hours in a few cases, due to the atom-typing procedure. The large-scale evaluation was performed with Maestro 8.0. No difference in terms of cavity detection performance was observed between both versions. However, the computation time was noticeably improved, down to about 2 min per structure.

Fpocket Full-Scale Evaluation. On the set of 5416 holo structures, fpocket was able to retrieve more than 75% of binding sites with a RO around 1. Similar to the other methods evaluated here, at an RO > 0.5 around 95% of all binding sites could be retrieved. As shown on Figure 5, although being a pure geometry-based method, fpocket has a very similar ranking performance to SiteMap, ranking 78% of all found binding sites on the top rank and 97% within the top 5 ranks.

Considering the set of 9900 apo structures, fpocket managed to correctly identify 82% of all binding sites (Figure 4), ranking 42% of those on rank 1 and 86% among the top 5 ranks (Figure 5). Here again, a drop in ranking performance can be seen compared to the results obtained on holo structures.

Calculation time with fpocket varied between 1 to 3 s per structure.

PocketFinder Evaluation. The results published by An et al. have been taken directly to compare PocketFinder to the other three methods evaluated here. As the initial data set published by An et al. has been slightly modified, these modifications were taken into account in the results presented here.

PocketFinder is able to retrieve around 55% of all holo binding sites with a RO close to 1 and 97% of all binding sites with a RO > 0.5. Although solely volume based, the ranking performance of PocketFinder appears to be generally

better than the other methods, allowing retrieval of 82% of found binding sites on the top rank and 99% on the top five ranks.

Next to the RO for measuring accuracy of pocket prediction, An et al. introduced in their evaluation two other measures extending the assessment of accuracy. The first is the ratio between the binding site and the ligand volumes, and the second is the ratio of the predicted binding patch with respect to the whole protein surface. Neither of these two criteria is used in the present study but rather the MO criterion. As all the results for PocketFinder were directly taken from the publication of An et al., no further calculation of the MO criterion was possible.

On the apo structure data set, PocketFinder identifies about 95% of all binding sites. Also the ranking performance of PocketFinder appears to be satisfying, ranking 80% of all found binding sites on the top rank and nearly 100% within the to 5 ranks.

Comparison between SiteFinder, SiteMap, PocketFinder, and fpocket. Evaluation results for all four algorithms on holo structures are summarized in Figures 4−6. Regarding the capacity of all algorithms to actually find the known binding site, the difference between them is rather small. SiteFinder (default parameters), fpocket, and SiteMap predict around 95% of the known binding sites with a RO above 0.5 (Figure 4). This result is rather interesting given that some of the methods are based on rather different methodologies. A significant improvement in predictive power can be observed for SiteFinder using optimized parameters.

Big differences in performance between SiteFinder (default parameters) and all other algorithms and parameters sets can be seen with respect to ranking of the binding sites. Figure 5 illustrates clearly that the predictiveness of SiteFinder is around 5% lower than the one of the other methods. The results obtained for fpocket, SiteMap, and SiteFinder (optimized parameters) are very similar. PocketFinder shows a slightly better ranking performance, although the ranking is simply based on the pocket volume.

Having observed the big performance increase between the default parameter set and optimized parameter set for SiteFinder as illustrated in Figure 4, one can state nevertheless that the accuracy of the prediction is influenced by changes of these parameters. As shown in Figure 6, the MO for SiteFinder (optimized parameter) is clearly lower than for all other methods. Comparing SiteFinder (default parameters) with fpocket and SiteMap another interesting observation could be made. Although both SiteFinder and fpocket are pure geometry-based approaches, they (especially fpocket) appear to predict binding sites slightly more accurately than SiteMap. This concurs with the smaller RO values obtained for fpocket and SiteFinder around RO = 1, indicating that both geometry algorithms generally produce smaller pockets than SiteMap.

Using all algorithms with the standard parameters, SiteMap clearly outperforms SiteFinder, fpocket, and PocketFinder regarding full coverage of the actual binding sites. Taking a half-covered binding site as correctly identified (as considered here), all methods perform well with a comparable level of predictiveness. Regarding ranking performance, only SiteFinder shows a clearly lower predictive power.

Considering accuracy of prediction, fpocket appears to propose pockets with better MO compared to all other methods.

The evaluation of predictive power of all methods on apo structures allows for the identification of further differences between the four algorithms. First of all, it should be pointed out that PocketFinder results were again taken directly from An et al. As the study on apo structures involves notably a step of structural alignment, it could be a source of variations in pocket definitions using the ligand present in the superimposed holo structure. In the present study, PyMOL's align function was applied on the chains known to hold the apo and holo binding sites. As the structural alignment procedure used by An et al. was not specified in the paper, we simply assumed that the method employed here produced comparable results. However, at least for fpocket, SiteMap, and SiteFinder, the very same protocol was applied, allowing a straightforward comparison between those algorithms.

Bearing these limitations in mind, one can observe that PocketFinder performs better than all other methods regarding accurate prediction of the binding site and the ranking. However, one should bear in mind that the comparison between the other methods and PocketFinder could be skewed. Among the other three methods SiteFinder with optimized parameters performs best, with the caveat that this comes at the cost of reduced accuracy and the prediction of very large binding sites. Among the algorithms with default parameters, fpocket performs best regarding accuracy of binding site prediction, while SiteMap clearly outperforms the other methods regarding ranking of binding sites.

DISCUSSION

A large-scale evaluation of four pocket prediction algorithms, SiteFinder, SiteMap, fpocket, and PocketFinder was performed. All algorithms were able to correctly predict binding sites in almost all proteins for the holo structures. The algorithm with default parameters that allowed the most binding sites with a high RO to be retrieved is SiteMap. By optimizing search parameters, SiteFinder outperforms all other methods at the cost of producing very big binding sites. Although this has no obvious primary sense, such a very comprehensive binding site detection can prove useful in cases where a pocket database is established for comparison of subpockets against other pockets/subpockets. SiteFinder's parameter set allows the construction of a representative collection of cavities containing entire binding sites, that is a cavity database. When exploiting such a database, one must bear in mind that the potential ligands may be smaller than the actual cavity.

Interestingly, SiteFinder using optimized parameters, SiteMap and fpocket show a surprisingly similar performance in ranking binding sites, although all three methods use completely different approaches. Solely SiteFinder (default parameters) appears to exhibit lower ranking performance compared to the latter. Another surprising finding is that both geometry-based methods (using default parameters) tend to produce more accurate binding sites than SiteMap, indicating that SiteMap predicts slightly bigger pockets than SiteFinder and fpocket.

A major performance drop was observed for predictions on apo structures. Assuming that a straightforward compari-

son between PocketFinder and the other algorithms is possible, PocketFinder outperforms all other algorithms. For the other three algorithms, using default parameters, fpocket performs best on accurate binding site prediction, while SiteMap performs best on ranking. Here again, both geometry-based algorithms appear to produce more accurate binding sites (cf., Figure 6). This finding, and the results shown in Figure 4, are at odds with the general idea that geometry-based criteria can only identify well-defined pockets as, for example, mentioned by T. Halgren.[30]

Based on our results it can be postulated that the concave curvature or high degree of burial is a common hallmark of all protein surface patches binding small molecules tightly. Compared to macromolecular interactions, this burial appears to be a necessity to provide sufficient shielding from solvent for a stable interaction to be possible. Furthermore, this will increase the number of contacts that a small molecule can make with the macromolecule and will therefore increase its binding efficiency at this site compared to locations on the macromolecule. Also, one can expect the local water structure to be more ordered within a small, concave, pocket. Release of these molecules into the bulk water upon binding of the ligand provides an entropic gain. Overall, relatively simple geometric rules are sufficient to account for these characteristics, and therefore, a corresponding algorithm can perform well at predicting and ranking binding sites.

Importantly, the herein used data set is not restricted with respect to the characteristics of the ligand molecules. Thus binding of physicochemically different small molecules (sugars, drugs, pro-drugs, etc.) requires generally a concave, solvent-shielded, portion of the protein surface.

Looking at the results published for the Cheng et al. data set[34] regarding the druggability of binding sites, it appears that there is a correlation between size as well as hydrophobicity and druggability (larger binding site, and increased hydrophobicity favoring druggability). Although the algorithm in SiteFinder does not explicitly target druggability, it is apparent that the way the alpha spheres are calculated and used for scoring implies as well that binding sites are ranked with respect to size and hydrophobicity. These characteristics, common to the different algorithms, also correspond to chemical intuition and to the trends very often observed in medicinal chemistry optimization programs, where larger and more hydrophobic molecules tend to have higher affinity. Nevertheless, optimizing molecules solely according to this criterion must be treated with caution because other properties, such as physicochemical and ADMET properties, tend to deteriorate at the same time.

Practical Considerations for Creating a Pocket Database. Both geometry-based algorithms have some inherent advantages over energy-based pocket prediction methods. First, the calculation time is about 90 times faster. Second, both geometry algorithms are robust against structural variations or missing atoms/residues that can occur in PDB files, as no atom-typing step and adding of H-atoms needs to be performed. Care must be taken, however, that the missing atoms or residues do not have an effect on the binding sites that are to be detected, which can be the case with geometry-based methods. In case of a detailed study of a system, where protonation states of all side chains in and around the binding site are known, energy-based pocket predictions can be very useful. However, this type of assignment on a high-throughput level is not realistic. If the task is the creation of a pocket database for the whole PDB or a large in-house databases, then geometry-based methods have a clear advantage.

In terms of userfriendliness, working environment, and informatics skills required, the algorithms cater to different tastes. Regarding the working environment, SiteFinder can be used through the very powerful SVL programming language available in the MOE or within the GUI itself. For SiteMap, the user can interact with the software through the graphical user interface available in Maestro or through the command line. Also, Maestro allows accessing molecular information through a Python-based API. However, for both of these algorithms a certain amount of effort in programming and automation has to be spent to adapt them for the creation of a pocket database, while working with SVL appears to be more straightforward, although it requires some knowledge of the SVL programming language. A very convenient algorithm for creation of a putative pocket database creation is fpocket. In essence it is standalone C code executable, and given that it is command line driven, it makes extraction of pocket scores, ranks, and descriptors very easy via a few command line flags. This information can then be organized using the tools the user prefers and is most comfortable with and not a preimposed working environment that the user has to adapt to.

CONCLUSIONS

In general, the binding site detection algorithms considered in this study exhibit a very good performance. Over 95% of all binding sites are retrieved within the 5 best ranked binding pockets. Considering the trade-off between speed and quality of the results, geometry-based methods like SiteFinder (using optimized parameters) or fpocket appear to be slightly more appropriate for creating a large cavity database for further use by cavity comparison algorithms.

Regarding SiteMap, it would be desirable to improve the treatment of structures with missing atoms or residues, in particular, if it is intended to be used for a systematic study or the preparation of a cavity database. This has been partly accomplished by the prepwizard program provided by Schrödinger. Nevertheless, it adds another intermediate step before pocket prediction using SiteMap.

Given that during the cavity detection process already a number of descriptors for the binding sites are calculated, possible extensions of the binding site algorithms would be the inclusion of a druggability score. Such druggability scores can be based on relatively simple descriptors,[34] as published in recent papers on SiteMap[30] and on fpocket,[32] and provide additional valuable information for the user. In the same vein, the calculated characteristics of a given binding site could be used in order to choose those molecules that should be screened first against the site of interest. Here one could imagine translating the binding site characteristics into a query for shape-based and/or pharmacophore-based screening, in order to identify molecules that are complementary to the binding site characteristics.

Combining the binding site detection algorithms with a binding site comparison tool would allow for the prediction of ligands likely to bind to a new binding site if the ligands for a similar binding site are known.

With the advent of systems biology and pathway- or network-based drug discovery,[35] binding site detection and characterization may gain additional importance, in particular if the idea to interfere at the same time with several targets—representing key players in a pathway or network—gains traction. In this case it might be necessary to find and compare binding sites on several different proteins in a given network with a view to identifying ligands that can bind to these proteins simultaneously, albeit with lower affinity. Binding site detection algorithms that are fast and efficient could prove invaluable for such an undertaking.

Note Added after ASAP Publication. This paper was published ASAP on September 9, 2010 with minor text errors and a corrected version was published on November 12, 2010. The version published ASAP on November 12, 2010 had an error in the optimized and default minrad values. The corrected version was published ASAP on November 17, 2010.

Supporting Information Available: Two files are provided, and both are in csv format and contain the PDB codes of the holo structures (LP_SETFinal.csv) used in this evaluation with corresponding ligand accessions. Another file named UP_SETFinal.csv contains the apo data set used in this study. This information is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *2*, 137–145.
(2) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *2*, 279–294.
(3) Alexandrov, N. N. SARFing the PDB. *Protein Eng.* **1996**, *9*, 727–732.
(4) Gibrat, J. F.; Madej, T.; Bryant, S. H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **1996**, *3*, 377–385.
(5) Madej, T.; Gibrat, J. F.; Bryant, S. H. Threading a database of protein cores. *Proteins* **1995**, *3*, 356–369.
(6) Holm, L.; Sander, C. Dictionary of recurrent domains in protein structures. *Proteins* **1998**, *1*, 88–96.
(7) Holm, L.; Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **1993**, *1*, 123–138.
(8) Ye, Y.; Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **2003**, ii246–ii255.
(9) Ye, Y.; Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.* **2004**, W582–W585.
(10) Brady, G. P., Jr.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *4*, 383–401, 0920−654; 0920−654.
(11) Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **1995**, *5*, 323–30, 307−8.
(12) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Modell.* **1997**, *6*, 359–63, 389.
(13) Huang, B.; Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, 19.
(14) Levitt, D. G.; Banaszak, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **1992**, *4*, 229–234.
(15) Peters, K. P.; Fauck, J.; Frommel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **1996**, *1*, 201–213.
(16) Binkowski, T. A.; Adamian, L.; Liang, J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.* **2003**, *2*, 505–526.
(17) Edelsbrunner, H.; Facello, M.; Fu, P.; Liang, J. Measuring proteins and voids in proteins. Proceedings of 28th Hawaii International Conference on System Science, Hawaii, January 4−7, 1995; IEEE: Piscataway, NJ,1995; pp 256264.
(18) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *9*, 1884–1897.
(19) Le Guilloux, V.; Schmidtke, P.; Tufferry, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **2009**, *1*, 168–178.
(20) Labute, P.; Santavy, M. Locating Binding Sites in Protein Structures. Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2001; http://www.chemcomp.com/journal/sitefind.htm. Accessed on June 30, 2010.
(21) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics* **2005**, *6*, 752–761.
(22) Cummins, P. L.; Titmuss, S. J.; Jayatilaka, D.; Bliznyuk, A. A.; Rendell, A. P.; Gready, J. E. Comparison of semiempirical and ab initio QM decomposition analyses for the interaction energy between molecules. *Chem. Phys. Lett.* **2002**, *3−4*, 245–251.
(23) Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L., III; Vajda, S. Algorithms for computational solvent mapping of proteins. *Proteins* **2003**, *3*, 340–351.
(24) Silberstein, M.; Dennis, S.; Brown, L.; Kortvelyesi, T.; Clodfelter, K.; Vajda, S. Identification of Substrate Binding Sites in Enzymes by Computational Solvent Mapping. *J. Mol. Biol.* **2003**, *5*, 1095–1113.
(25) Nettles, J. H.; Jenkins, J. L.; Williams, C.; Clark, A. M.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Flexible 3D pharmacophores as descriptors of dynamic biological space. *J. Mol. Graph. Modell.* **2007**, *3*, 622–633.
(26) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *6*, 449–462.
(27) Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins. *J. Mol. Biol.* **1999**, *4*, 1093–1108.
(28) Chemical Computing, G. I. Molecular Operating Environment.
(29) Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* **2007**, *2*, 146–148.
(30) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, 1549–960.
(31) *Maestro*, version 8.0; Schrödinger L.L.C.: New York, 2009.
(32) Schmidtke, P.; Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **2010**, *53*, 5858−5867.
(33) *R*, version 2.10.1; R Development Core Team: Vienna, Austria, 2009.
(34) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *1*, 71–75.
(35) Davis, J. C.; Furstenthal, L.; Desai, A. A.; Norris, T.; Sutaria, S.; Fleming, E.; Ma, P. The microeconomics of personalized medicine: today's challenge and tomorrow's promise. *Nat. Rev. Drug Discovery* **2009**, *4*, 279–286.

CI1000289

Structural Plasticity and Functional Implications of Internal Cavities in Distal Mutants of Type 1 Non-Symbiotic Hemoglobin AHb1 from Arabidopsis thaliana

Serena Faggiano, Stefania Abbruzzetti, Francesca Spyrakis, Elena Grandi, Cristiano Viappiani, Stefano Bruno, Andrea Mozzarelli, Pietro Cozzini, Alessandra Astegno, Paola Dominici, Silvia Brogioni, Alessandro Feis, Giulietta Smulevich, Oliver Carrillo, Peter Schmidtke, Axel Bidon-Chanal, and F. Javier Luque

Journal of Physical Chemistry B, 2009, 113

Structural Plasticity and Functional Implications of Internal Cavities in Distal Mutants of Type 1 Non-Symbiotic Hemoglobin AHb1 from *Arabidopsis thaliana*

Serena Faggiano,[†] Stefania Abbruzzetti,[‡] Francesca Spyrakis,[§,∥] Elena Grandi,[‡] Cristiano Viappiani,*[,‡] Stefano Bruno,[†] Andrea Mozzarelli,[†,∥] Pietro Cozzini,[§,∥] Alessandra Astegno,[⊥] Paola Dominici,[⊥] Silvia Brogioni,[#] Alessandro Feis,[#] Giulietta Smulevich,[#] Oliver Carrillo,[▽] Peter Schmidtke,[○] Axel Bidon-Chanal,[○] and F. Javier Luque[○]

Dipartimento di Biochimica e Biologia Molecolare, Università degli Studi di Parma, Parma, Italy, Dipartimento di Fisica, Università degli Studi di Parma, NEST CNR-INFM, Parma, Italy, Dipartimento di Chimica Generale ed Inorganica, Chimica Analitica, Chimica Fisica, Università degli Studi di Parma, Parma, Italy, INBB, Istituto Nazionale Biostrutture e Biosistemi, Consorzio Interuniversitario, Viale medaglie d'Oro 305, 00136 Rome, Italy, Dipartimento di Biotecnologie, Università degli Studi di Verona, Verona, Italy, Dipartimento di Chimica, Università degli Studi di Firenze, Sesto Fiorentino (FI), Italy, Department de Bioquímica i Biologia Molecular, Facultat de Biologia, Universitat de Barcelona, Barcelona 08028, Spain, and Departament de Fisicoquimica and Institut de Biomedicina (IBUB), Facultat de Farmacia, Universitat de Barcelona, 08028, Barcelona, Spain

Received: August 3, 2009; Revised Manuscript Received: October 19, 2009

The increasing number of nonsymbiotic plant hemoglobins discovered in genomic studies in the past decade raises intriguing questions about their physiological role. Among them, the nonsymbiotic hemoglobin AHb1 from *Arabidopsis thaliana* deserves particular attention, as it combines an extremely high oxygen affinity with an internal hexacoordination of the distal histidine HisE7 to the heme iron in the absence of exogenous ligands. In order to gain insight into the structure−function relationships of the protein, the ligand binding properties of mutants of two conserved residues of the distal cavity, HisE7 → Leu and PheB10 → Leu, were investigated by experimental and computational studies and compared to results determined for the wild type (wt) protein. The Fe^{2+}-deoxy HisE7 → Leu mutant exists, as expected, in the pentacoordinated form, while a mixture of penta- and hexacoordinated forms is found for the PheB10 → Leu mutant, with an equilibrium shifted toward the pentacoordinated form with respect to the wt protein. Spectroscopic studies of the complexes of CO and CN^- with AHb1 and its mutants show a subtle interplay of steric and electrostatic effects by distal residues on the ligand binding to the heme. Moreover, stopped-flow and flash photolysis experiments reveal substantial kinetic differences triggered by those mutations, which are particularly manifested in the enhanced geminate rebinding and bimolecular association rate. These findings are discussed in light of the drastic alterations found by molecular dynamics simulations in the nature and distribution of internal cavities in the protein matrix of the mutants, revealing an extremely large sensitivity of the protein structure to changes in distal HisE7 and PheB10 residues. Overall, data are consistent with the putative NO-dioxygenase activity attributed to AHb1.

Introduction

Class 1 nonsymbiotic hemoglobins (nsHbs) are found in a variety of plants, including rice,[1] barley,[2] maize,[3] tomato,[4] and *Arabidopsis thaliana*.[5,6] These globins exhibit a very high O_2 affinity mainly due to a low O_2 dissociation constant, which means that O_2 is stabilized after binding.[7] This finding suggests that they are are unlikely to be, *in vivo*, O_2 sensors, carriers, or storage molecules, since they would remain oxygenated even at very low O_2 concentrations. Moreover, their high redox potential also argues against a functional role in electron transport.[8,9] Several environmental factors induce expression of type 1 nsHbs. In particular, expression of AHb1 in *Arabidopsis thaliana* is induced by low levels of O_2[5] and exposure to nitrate[10] in both roots and rosette leaves, this latter response resembling the behavior of barley Hb.[11] Recently, AHb1 has been suggested to participate in NO detoxification by acting as a NO scavenger, thus reducing NO levels under hypoxic stress.[12−14]

Even though no X-ray crystallographic structure is available for AHb1 yet, it is known that Fe^{2+}-deoxy AHb1 is partly hexacoordinated, with the distal histidine, His69(E7) (hereafter simply denoted as HisE7), occupying the sixth coordination of heme iron in around 60% of the molecules.[15] Moreover, resonance Raman studies of CO complexes of AHb1 support the involvement of polar or hydrogen-bonding interactions in ligand stabilization, which could arise from HisE7 and Phe36(B10) (denoted as PheB10 in the following) residues.[15] Similar interactions have also been noticed in rice type 1 nsHb,[16]

* Corresponding author. Mailing address: Dipartimento di Fisica, Università degli Studi di Parma, viale G.P. Usberti 7/A, 43100 Parma, Italy. Phone: +390521905208. Fax: +390521905223. E-mail: cristiano.viappiani@fis.unipr.it.
[†] Dipartimento di Biochimica e Biologia Molecolare, Università degli Studi di Parma.
[‡] Dipartimento di Fisica, Università degli Studi di Parma.
[§] Dipartimento di Chimica Generale ed Inorganica, Chimica Analitica, Chimica Fisica, Università degli Studi di Parma.
[∥] INBB, Istituto Nazionale Biostrutture e Biosistemi.
[⊥] Università degli Studi di Verona.
[#] Università degli Studi di Firenze.
[▽] Department de Bioquímica i Biologia Molecular, Universitat de Barcelona.
[○] Departament de Fisicoquimica and Institut de Biomedicina (IBUB), Universitat de Barcelona.

10.1021/jp9074477 CCC: $40.75
Published on Web 11/12/2009

which is partially associated as a homodimer under physiological conditions and crystallizes as a dimer.[17] In this protein, the distal histidine hexacoordination is accompanied by a bend in the E-helix, the lack of a D-helix, and a disordered CD-region.[18] Although PheB10 pushes the distal histidine toward the heme propionates, HisE7 is close enough to the Fe atom to form a strong hydrogen bond with bound ligands, as judged from the low CO stretching frequency.[16] Moreover, mutations of PheB10 increase the binding constant of HisE7 to the heme iron and the auto-oxidation rate, suggesting that this residue is critical in affecting hexacoordination and promoting stabilization of bound exogenous ligands by HisE7.[16] Though the conservation of PheB10 might be important for stabilization of the oxy−ferrous complex,[16] this fact is of little help in assessing the functional role of nsHbs, as such stabilization is in fact needed for both O_2 transport and NO scavenging. However, the slower autoxidation rate and lower hexacoordination affinity constant conferred by PheB10 are not shared by other hexacoordinated Hbs such as neuroglobin, cytoglobin, or *Synechocystis* Hb.[3]

In order to establish the functional role of the two key amino acids, HisE7 and PheB10, mutants of AHb1 were investigated by combining experimental and computational studies. Electronic absorption spectroscopy and resonance Raman were used to explore the influence of these mutations on the distal cavity. Moreover, stopped-flow and laser flash photolysis were used to characterize the reactivity with CO, while steady state and time-resolved spectroscopies were performed to evaluate the effect on hexacoordination stability and ligand binding. Finally, information about the structural changes promoted by those mutations in the internal cavities was gained from extended molecular dynamics simulations. The results support the crucial role played by HisE7 and PheB10 not only in assisting ligand binding but also in modulating ligand migration pathways through the protein matrix. Altogether, these findings shed light onto the functional properties of AHb1, which is a prerequisite to understand its physiological role as a NO scavenger in *Arabidopsis thaliana* and, more generally, of nsHbs in plants.

Materials and Methods

Recombinant Protein Production and Purification. The cDNA encoding AHb1 was inserted into pET11a (Novagen) and used to transform *E. coli* BL21(DE3). The expression of recombinant proteins was carried out in the presence of 30 μM hemine chloride at 24 °C. Recombinant AHb1 was purified by chromatography on a Q-Sepharose Fast Flow (GE Healthcare) column eluted with a 100 mM Tris buffer at pH 7.2. The protein was then loaded on a Q-Sepharose High Performance (GE Healthcare) column, and a linear gradient of NaCl from 0 to 0.1 M in 20 mM Tris buffer at pH 8.5 was used for elution. The coding sequence for AHb1, cloned in the vector pGEM-T Easy (Promega), was used as a template to introduce mutations Phe36B10 → Leu and His69E7 → Leu by means of a QuikChange II mutagenesis kit (Stratagene) following the manufacturer's protocol. The mutants were expressed and purified as described for the wt protein. The yield of mutant proteins was comparable with that of wt AHb1.

Sample Preparation. The deoxy wt AHb1 and its mutants were prepared for stopped-flow experiments by diluting the concentrated stock of proteins with a deoxygenated buffer containing 100 mM sodium phosphate, 1 mM EDTA pH 7.0 to a final concentration of 20−30 μM. Sodium dithionite was added to a final concentration of 2 mM. The protein solution was mixed in the stopped-flow apparatus with the same deoxygenated buffer solution, containing 2 mM dithionite, equilibrated with nitrogen/CO mixtures of known CO partial pressure. For flash photolysis experiments, hemoglobin solutions were diluted in deoxygenated 100 mM sodium phosphate buffer, 1 mM EDTA, pH 7.0 to a final concentration ranging from 60 to 70 μM. Before the experiment, solutions were equilibrated with nitrogen/CO mixtures of known CO partial pressure and sodium dithionite was added to a final concentration of 2 mM.

SCHEME 1: Relevant Chemical Equilibria for the Reaction of CO with AHb1 or Its Mutants (Hb) to Form the CO Complex (HbCO)[a]

$$\begin{array}{ccccc} & & \left(Hb_p:CO\right)_1 & & \\ & & {\scriptstyle k_c}\uparrow\downarrow{\scriptstyle k_{-c}} & & \\ HbCO & \underset{k_{-1}}{\overset{h\upsilon}{\rightleftharpoons}} & \left(Hb_p:CO\right) & \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} & Hb_p + CO \\ & & & & {\scriptstyle k_{-h}}\uparrow\downarrow{\scriptstyle k_h} \\ & & & & Hb_h + CO \end{array}$$

[a] Penta- and hexacoordinated species were indicated by the suffix p and h, respectively. (Hb_p:CO) indicates primary docking sites with CO still inside the distal pocket, while $(Hb_p:CO)_1$ indicates a site in which the photodissociated ligand is docked into an internal hydrophobic cavity, accessible from the primary docking site.

CO complexes for RR spectroscopy were prepared by first flushing the protein solutions (30 μM) with nitrogen, then flushing with ^{12}CO (Rivoira) or adding ^{13}CO (FluoroChem), and finally adding dithionite (Fluka Chemicals) to reach a final 20 mM concentration. CN complexes were prepared by adding a few microliters of a diluted solution of potassium cyanide to the protein ferric form.

Resonance Raman. RR spectra were obtained at room temperature with excitation from the 413.1 nm line of a Kr^+ laser (Coherent). The backscattered light from a slowly rotating NMR tube was collected and focused into a triple spectrometer (consisting of two Acton Research SpectraPro 2300i and a SpectraPro 2500i in the final stage with a 1800 or 3600 grooves/nm grating) working in the subtractive mode, equipped with a liquid nitrogen cooled CCD detector (Roper Scientific Princeton Instruments). The spectra were calibrated to an accuracy of 1 cm^{-1} for intense isolated bands with indene, acetone, acetonitrile, and CCl_4 as standards. The grating was 3600 grooves/mm, with a spectral resolution of 1 cm^{-1}, for the low frequency region and 1800 grooves/mm, with a spectral resolution of 3 cm^{-1}, for the high frequency region.

Kinetic Studies. Stopped-flow experiments at a single wavelength were carried out using a temperature-controlled apparatus (SX.18MV, Applied Photophysics) using a 75 W Xe lamp as the light source and a photomultiplier as the detector. The instrumental dead time was 1.5 ms.

Flash photolysis was carried out with the circularly polarized second harmonic (532 nm) of a Q-switched Nd, YAG laser, and a CW Xe arc lamp as a probe source. The transient absorbance traces were measured at 436 nm through a 0.25 m spectrograph with a 5 stage photomultiplier (the experimental setup has been described elsewhere).[15,19] The cuvette had an optical path length of 2 mm. The repetition rate was about 0.3 Hz to allow for full sample recovery between laser flashes.

Kinetic Analysis of CO Rebinding Kinetics. A maximum entropy method (MEM)[20,21] was used to retrieve model-independent lifetime distributions, as described previously.[19,22] We have followed the minimal model previously proposed for CO rebinding to AHb1 solutions[15,19,23] to describe the rebinding kinetics (see Scheme 1). The differential equations corresponding to Scheme 1 (*vide infra*) were solved numerically, and the rate constants appearing in the equilibrium were optimized to

obtain a best fit to the experimental data. Numerical solutions were determined by using the function ODE15s within Matlab 7.0 (The MathWorks, Inc.). Fitting of the numerical solution to experimental data (and optimization of microscopic rate constants) was obtained with a Matlab version of the optimization package Minuit (CERN).

In order to improve the retrieval of microscopic rate constants, data from flash photolysis and stopped flow monitored at the same temperature using different CO concentrations (at 0.1 and 1 atm) were simultaneously fitted. This global analysis was repeated at different temperatures between 10 and 40 °C. The activation parameters for the microscopic rate constants were determined from the resulting linear Eyring plots (see Table 2).

Molecular Dynamics. MD simulations were run for wt AHb1 in the hexacoordinated (6c) form and in the deoxy and oxy states of the pentacoordinated form (5c-deoxy and 5c-oxy, respectively). Additional simulations were also run for the 6c form of the PheB10 → Leu mutant and for the oxy state of 5c for the PheB10 → Leu and HisE7 → Leu mutants.

Starting models of 6c and 5c forms were built with the homology modeling program MODELLER (http://salilab.org/modeller/), using as templates the X-ray crystallographic structures of rice (PDB code 1D8U)[18] and barley (PDB code 2OIF)[24] hemoglobins, respectively (AHb1 shows a sequence identity close to 70% in both cases). The original CN group bound to the heme iron in the barley template was replaced with O_2. The orientation of side chains was adjusted using Sybyl tools (www.tripos.com), and the models were further checked with PROCHECK.[25] The standard protonation state at physiological pH was assigned to ionizable residues. Specifically, with the sole exceptions of proximal HisF8 and distal HisE7 residues, all histidines were maintained in the default tautomeric (N_ε−H) state. For HisF8, which occupies the fifth coordination position of the heme iron, the N_δ−H tautomeric form was used. In the 6c form, HisE7 was also modeled using this latter tautomer, but the N_ε−H form was used in the 5c-oxy state in order to allow the formation of a hydrogen bond with heme-bound O_2. Finally, in the 5c-deoxy state, HisE7 was modeled in the two tautomeric (HID, HIE) states.

The starting structures were immersed in a preequilibrated octahedral box of TIP3P[26] water molecules. The final systems contained around 7700 waters for 6c forms of AHb1 and its PheB10 → Leu mutant and 10100 waters for 5c forms of the wt protein and its PheB10 → Leu and HisE7 → Leu mutants. MD simulations were run using the parmm99 force field and the Amber-9 package.[27] The heme parameters were developed and tested in previous works.[28,29] The SHAKE algorithm was used to keep bonds involving hydrogen atoms at their equilibrium length, in conjunction with a 1 fs time step for the integration of Newton's equations. Trajectories were collected in the NPT (1 atm, 298 K) ensemble using periodic boundary conditions and Ewald sums (grid spacing of 1 Å) for long-range electrostatic interactions. The systems were minimized using a multistep protocol, involving first the adjustment of hydrogens, then the refinement of water molecules, and finally the minimization of the whole system. The equilibration was performed by heating from 100 to 298 K in four 100 ps steps at 150, 200, 250, and 298 K. Finally, for each simulated system, 50 ns production trajectories were run, collecting frames at 1 ps intervals.

The FPOCKET program[30] was used to detect internal cavities in 800 snapshots taken regularly from the last 40 ns of the trajectories. The identified cavities were superposed in time and space, and a density map was generated from this superposition. High density cavities correspond to stable cavities found during the trajectory, while low density cavities are transient or nearly nonexistent in the MD simulation. In addition, the most feasible ligand migration pathway was explored for the same set of snapshots using GRID-MD.[31] For each snapshot, the van der Waals and Poisson−Boltzmann electrostatic energies between the protein and a rigid ligand probe were calculated at every point of a grid (0.5 Å spacing) that enclosed the whole protein using a probe particle corresponding to a carbon atom with 1.8 Å radius. The calculation of the potential energy was extended to the whole set of snapshots by Boltzmann averaging of the energies computed at the grid points determined for each snapshot. Finally, the trajectories followed by the probe particle (at 298 K) were sampled using Brownian dynamics. For the 5c-oxy state, the probe was located around the sixth coordination position (after removing the heme-bound O_2 ligand). All calculations were performed in the *MareNostrum* supercomputer at the Barcelona Supercomputing Center.

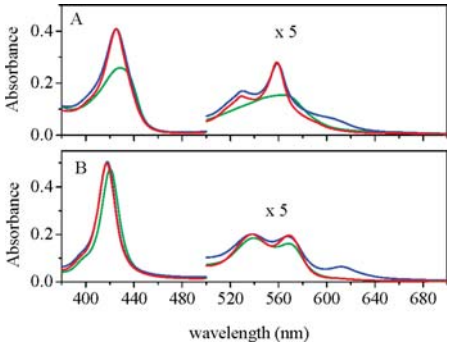


Figure 1. Absorbance spectra of the Fe^{2+}-deoxy (A) and CO-bound (B) forms of wt AHb1 (red line), PheB10 → Leu (blue line), and HisE7 → Leu (green line) mutants. $T = 20$ °C.

Results and Discussion

Electronic Absorption Spectroscopy. The absorption spectrum of Fe^{2+}-deoxy wt AHb1 reveals a mixture of a hexacoordinated and a pentacoordinated form, the former showing spectroscopic features similar to those of neuroglobin[32] and rice Hb1.[16] As expected, the mutation of distal histidine to leucine leads to the shift in the Soret band to 429 nm and to a broadband in the visible region centered at about 560 nm, features typical of a pure pentacoordinated high-spin state (5cHS; Figure 1A). In contrast, the absorption spectrum of the Fe^{2+}-deoxy form of mutant PheB10 → Leu (Figure 1A) shows peaks at 424 nm (Soret band) and at 558 and 529 nm (α and β bands, respectively).[15] A broadening of the red side of the Soret band and of the α and β bands with respect to the spectra of pure hexacoordinated globins indicate for this mutant, as for wt AHb1, an equilibrium between an hexacoordinated and a pentacoordinated form. Noteworthy, the 5cHS species is more populated in the mutated protein.

Using the absorbance spectra of Fe^{2+}-deoxy AHb2[15] and Fe^{2+}-deoxy HisE7 → Leu mutant as references for pure hexacoordinated low-spin (6cLS) and 5cHS species, respectively, the fraction of 5cHS species in Fe^{2+}-deoxy PheB10 → Leu is estimated to be around 63%, with an equilibrium constant of ∼0.59 (Supporting Information). For comparison, the equilibrium constant for wt AHb1 is ∼1.12 under the same experimental conditions (Supporting Information). Thus,

hexacoordination is apparently favored by PheB10, and removal of this residue weakens the tendency of HisE7 to bind the heme Fe. This finding is fairly surprising, as the results determined for several B10 mutants of rice Hb1 show the opposite trend. In particular, for rice type 1 nsHb, the ratio A_{555}/A_{540}, taken as an indicator of the degree of heme hexacoordination, showed that replacement of PheB10 with leucine favors the formation of a pure 6cLS species.[16]

For CO complexes of wt AHb1 and mutated proteins (Figure 1B), the Soret band (417 nm in the wt protein) is almost unaffected in the PheB10 → Leu mutant (418 nm). However, a clear red shift (420 nm) is found for the HisE7 → Leu mutant. Similarly, the peaks corresponding to α and β bands (569 and 538 nm in wt AHb1) shift to 568 and 538 nm in PheB10 → Leu and to 569 and 539 nm in HisE7 → Leu. Since CO complexes of heme proteins do not display absorption bands in the 600−700 nm range (due to the absence of charge transfer transitions for low-spin heme proteins in this region), a weak band at 610 nm in the spectrum of PheB10 → Leu is peculiar and likely reflects a small amount of an impurity (weak bands at this wavelength could often be seen in the spectra of CO complexes of heme proteins).[33,34]

Resonance Raman Spectroscopy. RR spectroscopy of Fe^{2+}-deoxy (data not shown) confirmed that the HisE7 → Leu mutation leads to a pure 5cHS species, while the PheB10 → Leu mutant preserves the partial hexacoordination observed in the wt protein,[15] though with a slight shift of the equilibrium toward the 5cHS species.

RR spectra of CO complexes of wt AHb1 and mutated proteins were examined to gain insight into the electrostatic interactions with distal cavity residues (Figure 2). For the wt protein, bands at 533 and 1923 cm^{-1} have been previously assigned to the $\nu_{(FeC)}$ and $\nu_{(CO)}$ modes, respectively, on the basis of $^{12}C/^{13}C$ isotopic substitution.[15] It is well established that $\nu_{(FeC)}$ and $\nu_{(CO)}$ frequencies are inversely correlated, owing to d_π electron back-donation from Fe to CO, and that high $\nu_{(FeC)}$ frequencies together with low $\nu_{(CO)}$ frequencies indicate a polar environment around the bound CO.[35] This is the case for wt AHb1, for which the distal histidine is proposed to strongly interact with CO through polar or hydrogen-bonding interactions. In contrast, single $\nu_{(FeC)}$ and $\nu_{(CO)}$ bands at 501 and 1964 cm^{-1} for the CO complex of mutant HisE7 → Leu indicate a reduced polarity in the distal cavity of the mutated protein,[36] thus confirming that the distal histidine is the residue that mainly provides the H-bond to the heme-bound CO in wt AHb1. The effect of the HisE7 → Leu mutation on the CO liganded species is similar to what was recently reported for the analogous mutation (His73 → Leu) in rice nsHb1.[16]

Two CO conformers are observed for the PheB10 → Leu mutant. The less populated conformer is characterized by $\nu_{(FeC)}$ and $\nu_{(CO)}$ at 493 and 1965 cm^{-1} (490 and 1921 cm^{-1} in the ^{13}CO complex, respectively), which are typical of CO complexes where polar interactions between CO and the protein matrix are minimal because of either the distance or low polarity of the surrounding residues.[36] However, the main effect of the PheB10 → Leu mutation is a shift of the $\nu_{(FeC)}$ frequency from 533 to 519 cm^{-1} (515 cm^{-1} in the ^{13}CO complex). Surprisingly, the $\nu_{(CO)}$ frequency stays unaltered at 1923 cm^{-1} (1880 cm^{-1} in the ^{13}CO complex) upon mutation, whereas an upshift would have been expected according to the Fe back-donation mechanism. A similar situation is found in the RR spectra of CO complexes of wt AHb1 and AHb2,[15] but conversely, the PheB10 → Leu mutation in rice Hb1 shifts $\nu_{(CO)}$ from 1926 to 1937 cm^{-1} ($\nu_{(FeC)}$ frequencies are not available for rice Hb1).[16]

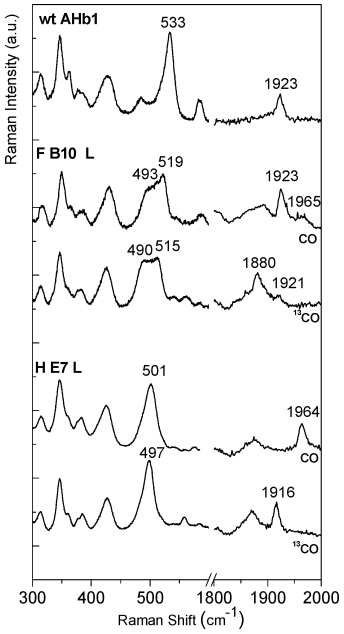


Figure 2. Top: RR spectrum of the CO complex of wt AHb1 (laser power was 2 mW, accumulation times were 140 and 40 min, respectively, for the low and high frequency regions). Middle: RR spectrum of ^{12}CO and ^{13}CO complexes of the PheB10 → Leu mutant (^{12}CO: laser power was 2 mW, accumulation times were 40 and 180 min, respectively, for the low and high frequency regions; ^{13}CO: laser power was 2 mW, accumulation times were 40 and 110 min, respectively, for the low and high frequency regions). Bottom: RR spectrum of ^{12}CO and ^{13}CO complexes of the HisE7 → Leu mutant (^{12}CO: laser power was 2 mW, accumulation times were 40 and 80 min, respectively, for the low and high frequency regions; ^{13}CO: laser power was 5 mW, accumulation times were 60 and 50 min, respectively, for the low and high frequency regions).

There are several examples of Phe mutations in the distal cavity of heme proteins where the Fe−CO electron back-donation, and consequently the vibrational frequencies, are influenced in a similar way. For example, substitution of Phe43 with valine in sperm whale myoglobin[37] shifts the main $\nu_{(CO)}$ peak from 1945 to 1954 cm^{-1}. The CO complex of *C. cinereus* peroxidase displays $\nu_{(CO)}$ and $\nu_{(FeC)}$ at 1930.5 and 519 cm^{-1}, while the corresponding bands in the Phe46 → Val mutant are at 1944 and 508 cm^{-1}.[38] All of these observations could reflect a decreased polarity of the CO environment upon Phe mutation (in fact, the side chain of Phe has a high quadrupole moment).[39] Alternatively, the decreased back-bonding in the PheB10 → Leu mutant could reflect a change in the position of the distal histidine due to the reduced steric hindrance of leucine, which could then be less favorable for an electrostatic interaction in the PheB10 → Leu mutant than in the wt protein. The interaction between PheB10 and HisE7 could also be the origin of the observed change in the 5c ⇔ 6c equilibrium in the PheB10 → Leu mutant (see results from electronic absorption spectroscopy and ligand binding kinetics).

Additional information can be obtained from the RR spectra of $Fe^{3+}-CN^-$ complexes. The $Fe^{3+}-CN^-$ adduct experiences much less backbonding than the $Fe^{2+}-CO$ complex, and the FeCN bond has mainly σ-bonding character. Accordingly, it is scarcely affected by the polarity of the surrounding medium and the spread of frequencies is small.[40−42] Nevertheless, the

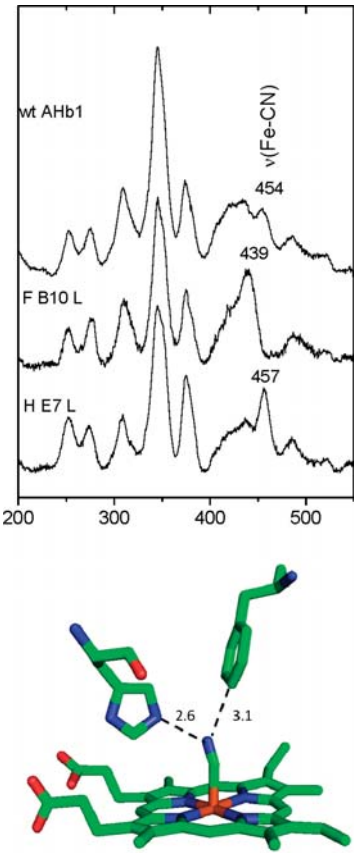


Figure 3. Top: RR spectra of the CN complexes of wt AHb1 (laser power 25 mW, accumulation time 80 min), the PheB10 → Leu mutant (laser power 12 mW, accumulation time 75 min), and the HisE7 → Leu mutant (laser power 12 mW, accumulation time 60 min). Bottom: Structure of the CN^- complex of barley nsHb1 (PDF entry 2OIF) showing the location of HisE7 and PheB10 residues (distances in Å).

$Fe^{3+}-CN^-$ adduct can be easily bent, and RR frequencies give valuable information about distal interactions of bent forms.[43] Thus, the $\nu_{(FeC)}$ band shifts to lower wavenumbers when sterically encumbering groups are in close contact with the bound CN^-. For example, $\nu_{(FeC)}$ of the CN^- complex of *C. eugametos* Hb[44] shifts from 440 cm^{-1} in the wt protein to 452 cm^{-1} when the distal residue Gln is replaced by Gly. Moreover, for a series of cyanoporphyrins with a progressively "tighter" distal side $\nu_{(FeC)}$ shifts from 451 cm^{-1} in the unhindered porphyrin to 445 cm^{-1} in the most hindered complex.[45]

The $\nu_{(FeC)}$ band of the CN^- complexes of wt AHb1 (454 cm^{-1}) and its PheB10 → Leu (439 cm^{-1}) and HisE7 → Leu (457 cm^{-1}) mutants was isolated from the shift promoted upon isotope substitution (Figure 3; see also Figures S2 and S3 in the Supporting Information). No additional isotope-sensitive bands were detected in the 200−600 cm^{-1} range. The small but significant difference in $\nu_{(FeC)}$ found between wt AHb1 and the HisE7 → Leu mutant reflects the steric hindrance in the distal side, which would deviate the FeC bond from the axis normal to the heme plane. In fact, inspection of the X-ray structure of the CN^- complex with barley nsHb (Figure 3)[24] suggests that the deviation from linearity of the FeCN group could be larger upon replacement of PheB10 by leucine, thus justifying the reduction in $\nu_{(FeC)}$ in the CN^- complex of PheB10 → Leu.

TABLE 1: Vibrational Frequencies (cm^{-1}) of the Fe−C−X Unit in CO and CN^- Complexes of wt AHb1 and Its Mutants

	ν(Fe−CO)	ν(CO)	ν(Fe−CN)
wt AHb1	533	1923	454
PheB10 → Leu	519	1923	439
HisE7 → Leu	501	1964	457

Comparison with the observations made on the CO complexes reveals the different effects of protein structural changes on the FeC vibrational frequencies in either case (Table 1). Tilting and bending of the FeCX unit has little impact in the case of FeCO,[35] whereas the effects of the environment polarity are evident and relatively well understood. The opposite seems to occur for the $FeCN^-$ adduct, although the relationship between spectral and structural properties is less clear. Substitution of HisE7 by leucine strongly influences the RR spectra of the CO complex due to loss of polarity but has a small effect on the spectra of the CN^- complex. On the other hand, the PheB10 → Leu mutation markedly changes the $\nu_{(FeC)}$ frequency of the CN^- complex, whereas its effects on the spectra of the CO complex are less dramatic than for the HisE7 → Leu mutation.

Ligand Binding Kinetics. Flash photolysis and stopped-flow experiments were carried out to determine the effect of mutations at PheB10 and HisE7 on the microscopic rates for CO binding and the competitive reaction of the endogenous His residue. Our previous stopped-flow experiments of CO binding to wt AHb1 had shown a biexponential kinetics due to binding to the 6cLS and 5cHS species (see also Figure 4 for a representative kinetic trace).[15] A biexponential relaxation is also found for the reaction with the PheB10 → Leu mutant (blue curve in Figure 4). In this case, however, a larger fraction of the reaction progress is lost in the instrumental dead time (1.5 ms) due to the high binding rate to the 5cHS species, in agreement with the shift of the equilibrium toward the 5c species evidenced in the RR and absorption spectra. Stopped-flow experiments for the HisE7 → Leu mutant show a single exponential kinetics (green curve in Figure 4), as expected for the existence of a single molecular species (5cHS) reacting with CO with a rate higher than that observed for the wt protein. In this case, most of the kinetics is lost within the instrumental dead time.

The CO rebinding kinetics to wt AHb1 and to the mutants after nanosecond laser photolysis at two temperatures are compared in Figure 5. In all cases, the increase in temperature reduces the amplitude of the geminate rebinding phase and

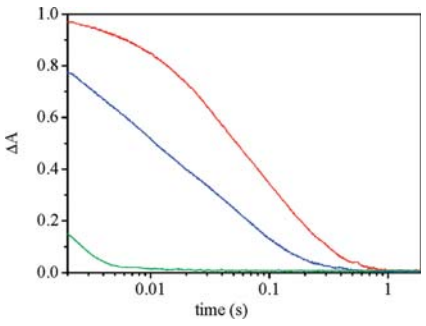


Figure 4. Stopped-flow CO binding kinetics to wt AHb1 (red), PheB10 → Leu AHb1 (blue), and HisE7 → Leu AHb1 (green) at 5 °C and 0.1 atm CO.

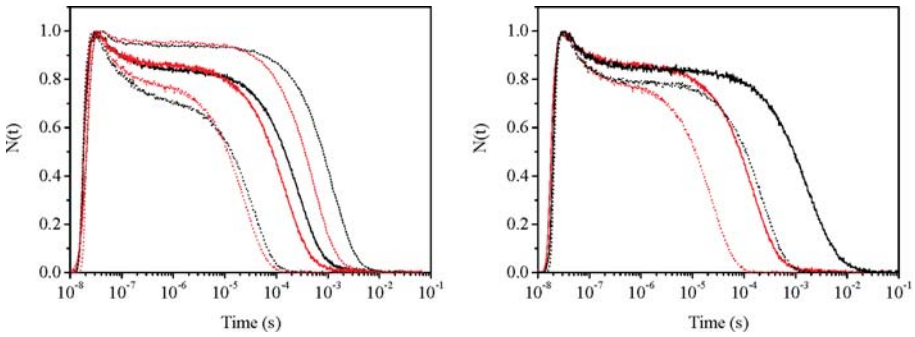


Figure 5. Left: Comparison between CO rebinding kinetics to wt AHb1 (short dashed), PheB10 → Leu (solid), and HisE7 → Leu (dotted) at 10 °C (black) and 30 °C (red). Solutions were equilibrated with 1 atm CO. The protein concentration was 60−70 μM. Right: Comparison between rebinding kinetics to AHb1 mutants PheB10 → Leu (60 μM, solid) and HisE7 → Leu (80 μM, dotted) at 30 °C for solutions equilibrated with 1 atm CO (red) and 0.1 atm CO (black).

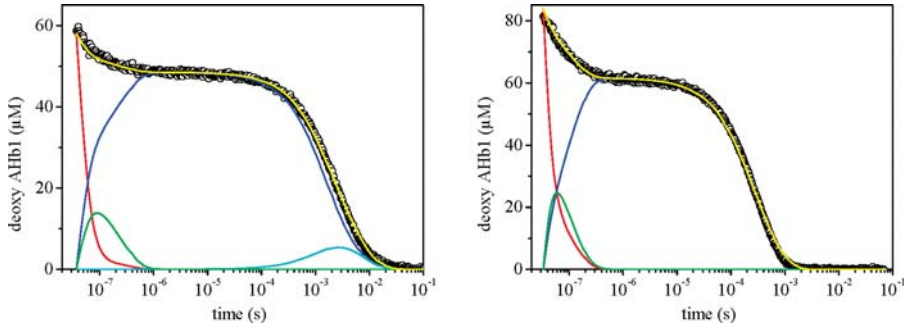


Figure 6. Analysis of the CO binding kinetics to the PheB10 → Leu (left) and HisE7 → Leu (right) mutants at 0.1 atm CO and 10 °C. The fit (yellow line) is superimposed to the experimental data (black open circles). The time course of relevant species in Scheme 1 is also shown: (Hb_p:CO), red; $(Hb_p{:}CO)_1$, green; Hb_p, blue; Hb_h, cyan.

enhances the apparent rate of the bimolecular phase. The amplitude of geminate rebinding to both mutants is larger than for the wt protein. The effect of temperature on this kinetic phase is modest for the PheB10 → Leu mutant, while the HisE7 → Leu mutant exhibits slightly higher temperature sensitivity. This finding may suggest that HisE7 is slightly more relevant than PheB10 in modulating the dynamics of the distal pocket, as far as the ligand rebinding is concerned. However, it might also occur that the HisE7 → Leu mutation exposes an exchange pathway with higher temperature sensitivity. Both mutations appear to disfavor ligand exit to the solvent phase, as demonstrated by the increased geminate recombination with respect to the wt protein. Another striking feature is the much higher apparent bimolecular rebinding rate in the mutants relative to the wt AHb1. This is not unexpected for the HisE7 → Leu mutant, which only populates the 5cHS form. However, a similar, though smaller effect is found for the PheB10 → Leu mutant, suggesting that ligation and detachment of the distal His occurs with higher rates/lower yield for this mutant. Finally, the CO concentration dependence of the rebinding kinetics (shown in Figure 5 for representative CO concentrations at 30 °C) shows a clear-cut separation between geminate and bimolecular rebinding, as previously noted for the wt protein.[19]

Following our previous work,[15] the rebinding kinetics has been interpreted in the framework of Scheme 1, which was well suited for the rebinding kinetics to wt AHb1 in solution. For the HisE7 → Leu mutant, the species Hb_h and (Hb_h:CO) were removed, since the heme is purely pentacoordinated in the absence of exogenous ligands. We have simultaneously analyzed the rebinding kinetics at different CO concentrations to improve the reliability of the retrieved parameters. This simultaneous analysis was performed at each temperature, and the activation parameters were determined from the temperature dependence of the rate constants. As an example, Figure 6 reports sample fits to the CO rebinding curves under selected conditions. A remarkably good agreement was obtained under all experimental conditions, thus showing that the model is robust and can be used to characterize the response of wt AHb1 to point mutations. The results are reported in Table 2 along with the activation enthalpies and entropies determined from Eyring plots of the microscopic rate constants.

The binding (k_b) and dissociation (k_{-b}) rates for the distal His are increased in the PheB10 → Leu mutant, with the equilibrium constant between 6cLS and 5cHS species ($K_H \approx 0.5$) being reversed with respect to the wt AHb1 ($K_H \approx 1.6$), leading to a lower fraction of 6cLS species (33%) in the mutant relative to the wt protein (61%). These findings are in keeping with the fractions estimated from absorbance spectra (Figure 1 and Supporting Information).

The increase in geminate recombination in the mutated proteins with respect to wt AHb1 seems to arise from different processes. While k_{-1} is scarcely influenced by the PheB10 → Leu mutation, it is slightly increased in the HisE7 → Leu mutant. The absence of thermal activation for k_{-1} in the temperature range examined here is shared by all of the samples. The exit rate k_2 is decreased ∼3.5-fold by the two mutations, with comparable effects on the activation enthalpy and entropy. The rate k_{-2} is dramatically increased (∼7-fold) for the HisE7 → Leu mutant, with strongly reduced activation parameters. A reduction in the activation parameters is also observed for the

TABLE 2: Microscopic Rate Constants and Activation[a] Enthalpies (kcal mol^{-1}) and Entropies (cal mol^{-1} K^{-1}) Determined for wt AHb1 and Mutated Proteins from the Global Fit of Flash Photolysis (at 1 and 0.1 atm CO) and Stopped-Flow (at 0.05 atm CO) Data at 20 °C

	wt[b]			PheB10 → Leu			HisE7 → Leu		
	k	$\Delta S^{\ddagger}$	$\Delta H^{\ddagger}$	k	$\Delta S^{\ddagger}$	$\Delta H^{\ddagger}$	K	$\Delta S^{\ddagger}$	$\Delta H^{\ddagger}$
k_{-1} (10^6 s^{-1})	5.13			5.2			8		
k_2 (10^7 s^{-1})	9	-12.9 ± 0.6	2.7 ± 0.1	2.6	-16 ± 3	2.4 ± 0.9	2.7	-14.8 ± 0.9	2.8 ± 0.3
k_{-2} (10^7 M^{-1} s^{-1})	2.26	26 ± 2	14.7 ± 0.6	3.1	4.8 ± 0.4	8.5 ± 0.1	16	4.8 ± 1.5	7.6 ± 0.4
k_b (s^{-1})	23.5	8 ± 4	18 ± 1	44.2	12 ± 5	18 ± 1			
k_{-b} (s^{-1})	14.5	12 ± 4	19 ± 1	79.6	18 ± 2	19.9 ± 0.6			
k_c (10^7 s^{-1})	2.07			1.48			3.5		
k_{-c} (10^7 s^{-1})	0.25			7.53			3.5		

[a] Activation enthalpies $\Delta H^{\ddagger}$ and entropies $\Delta S^{\ddagger}$ were estimated from Eyring plots for each rate constant k_i in the temperature range 10−40 °C, according to the equation $\ln(hk_i/k_BT) = \Delta S^{\ddagger}/R - \Delta H^{\ddagger}/RT$, where R is the gas constant, h is Planck's constant, and k_B is the Boltzmann constant. [b] Data from ref 19.

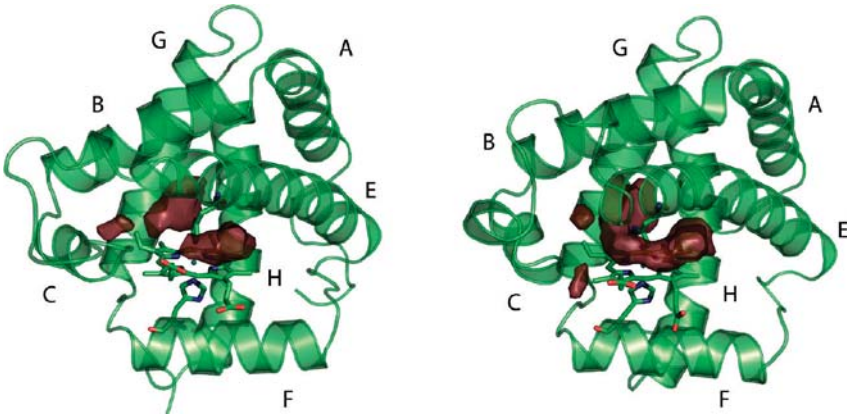


Figure 7. Representation of the cavities found from FPOCKET computations in the 6c state of AHb1 (left) and its PheB10 → Leu mutant (right).

PheB10 → Leu mutant, but the increase in the rate is not as large. Finally, it is worth noting the large difference found for the rebinding process from a secondary docking cavity to the primary distal cavity (k_{-c}), which is increased ∼30-fold and 14-fold for the PheB10 → Leu and HisE7 → Leu mutants, respectively, compared to wt AHb1.

Overall, the effects on the rate constants result in a substantial increase for k_{ON} ($\approx k_{-2}k_{-1}/k_{2+}k_{-1}$), which is 1.22×10^6 M^{-1} s^{-1} for the wt AHb1 at 20 °C, and becomes 5.17×10^6 M^{-1} s^{-1} for PheB10 → Leu and 3.66×10^7 M^{-1} s^{-1} for HisE7 → Leu. By comparison, the rate constant for CO binding to wt rice Hb is 7×10^6 M^{-1} s^{-1} and increases to 9×10^6 M^{-1} s^{-1} for the PheB10 → Leu mutant.[16]

Molecular Modeling. In order to examine the structural features of cavities or tunnels potentially implicated in the migration of diatomic ligands through the protein matrix, a series of MD simulations were performed for wt AHb1 and the mutants. In all cases, inspection of the profiles determined for time evolution of the potential energy and root-mean-square deviation (rmsd) along the trajectories supported the stability of the MD simulations and the validity of homology models (see Figure S4 in the Supporting Information). In particular, fluctuations in the rmsd, mainly due to changes in CD and EF loops, were found along the first 10 ns. Stable rmsd values ranging from 2.0 to 2.7 Å relative to the energy-minimized structure of the protein were determined for the rest of the trajectory (see the Supporting Information). Accordingly, the analysis of the simulations was limited to the snapshots sampled in the 10−50 ns region.

In the 6c state, the distal cavity (defined by residues Leu35, Leu36, Ile39, Phe50, Leu66, Ala70, Val73, Ala117, and Leu121) remains stable along the trajectories sampled for AHb1 and its PheB10 → Leu mutant. In the wt protein, two cavities are found above the heme surrounding the coordinated distal His, while those cavities are connected in the mutant as a consequence of the mutation of the distal Phe (Figure 7). Few water molecules were found in the interior of the heme cavity along the trajectory, a fact that supports the accessibility from the aqueous solvent, in agreement with the biexponential kinetic behavior found for the hexacoordinated forms (see above).

The nature and distribution of cavities in the 5c-deoxy form of wt AHb1 were examined considering the distal His in two tautomeric states (Figure 8). A distal cavity was constantly present along the trajectories sampled for the wt protein using the HIE and HID tautomeric forms of HisE7. Moreover, in the two cases, an additional small cavity lined by residues Cys77, Cys78, Ala81, Trp141, Ala144, Tyr145, and Leu148 was found in the protein matrix but without an apparent, permanent linkage with the distal cavity. On the other hand, the increased flexibility of the distal His leads to larger fluctuations of the dihedral angles of the side chain compared to the 6c state. In particular, a drastic conformational change was observed in the case of the HID tautomer after the first 15 ns of the trajectory (Figure 9; see Figure S5 in the Supporting Information). Noteworthy, such conformational rearrangement mimics the opening of the distal histidine gate observed in carbonmonoxymyoglobin,[46] which would thus facilitate the formation of a path that connects the distal site with the exterior.

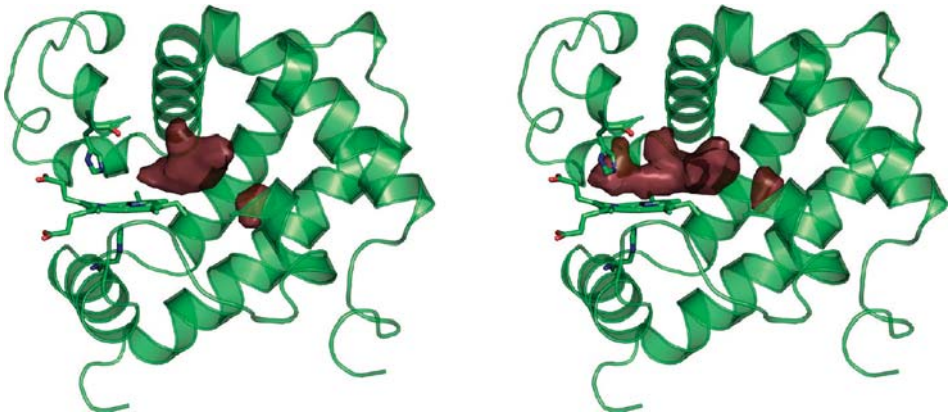


Figure 8. Representation of the cavities found from FPOCKET computations for the 5c-deoxy state of wt AHb1 considering HE7 in its HIE (left) and HID (right) tautomeric states.

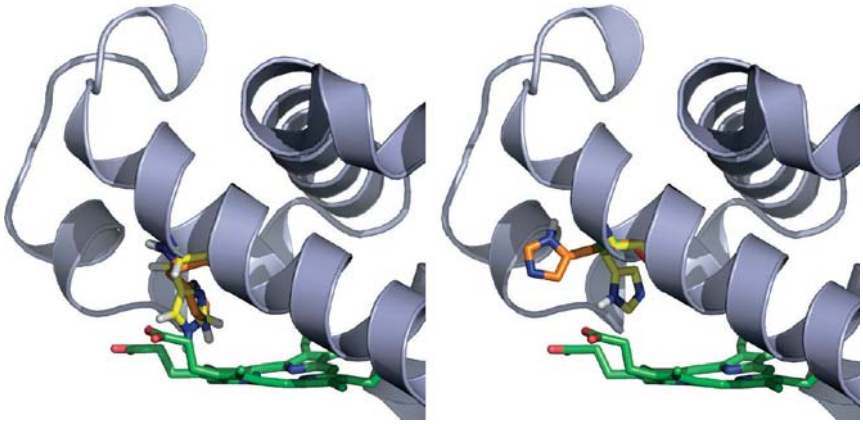


Figure 9. Representation of the orientations of the distal His in snapshots taken at 10 and 40 ns in the simulations of wt AHb1 in the 5c-deoxy state considering HIE (left) and HID (right) tautomeric forms.

Distinct trends concerning the shape and volume of the internal cavities are found in the 5c-oxy state of wt AHb1. Thus, FPOCKET results show the formation of a secondary cavity delineated by residues Cys77, Cys78, Ser80, Ala81, Leu84, Val90, Trp141, Ala144, His147, and Leu148, which is transiently accessible to the bulk solvent through rearrangements of the side chain of His147 (Figure 10). In fact, accessibility was confirmed by the presence of two water molecules along the simulation. More importantly, there is a direct connection between the distal cavity and the secondary one, which seems to be modulated by residues Cys77, Cys78, Leu121, and particularly by the side chain of Tyr145. Thus, in the 6c- and 5c-deoxy states, Tyr145 exhibits little fluctuations, as the orientation of the side chain is largely fixed by a hydrogen bond with the carbonyl group of Phe114 (average (Phe114)O···O(Tyr145) distance of 2.9 Å for wt AHb1 in the 6c state, and between 3.1 and 3.3 Å in the 5c-deoxy form). In contrast, this interaction is more labile in the 5c-oxy form (average distance: 5.24 Å), and breaking of the Phe114−Tyr145 interaction is sometimes associated with the formation of a hydrogen bond between Tyr145 and Cys77 (see Figure S6 in the Supporting Information). Noteworthy, the formation of a transient pathway in the 5c-oxy state of wt AHb1 is supported by GRID-MD computations, which confirmed the feasibility of the channel for ligand migration (see Figure 10). Overall, these findings are in agreement with the results coming from laser flash photolysis experiments, which showed a very low fraction of geminate rebinding for wt AHb1, as expected from the presence of a channel connecting the distal heme cavity with the exterior.[15]

The analysis of the trajectory sampled for the oxy state of the PheB10 → Leu mutant reveals some analogies with wt AHb1, but also some differential trends. First, the interaction between Phe114 and Tyr145, which was maintained along the 6c PheB10 → Leu trajectory (average distance: 2.83 Å), is broken, and leads to temporary hydrogen-bond interactions between Tyr145 and Cys78 (see Figure S7 in the Supporting Information). These changes not only affect the size and shape of the distal cavity but also lead to the formation of new cavities. In particular, a large cavity lined by residues Cys77, Ser80, Val92, Leu100, Phe114, Ala117, Tyr145, Leu148, and Ile152 is formed under the heme (Figure 10), which is transiently connected to the distal cavity. Even though this new cavity partially overlaps with the secondary cavity found in oxy wt AHb1, it is enclosed in the interior of the protein and no direct connection with the bulk solvent was detected. Noteworthy, the presence of this alternative docking site agrees with the significant increase in geminate rebinding observed for the PheB10 → Leu mutant, since small ligands could be trapped in the internal cavities close to the heme and then released into the distal pocket.

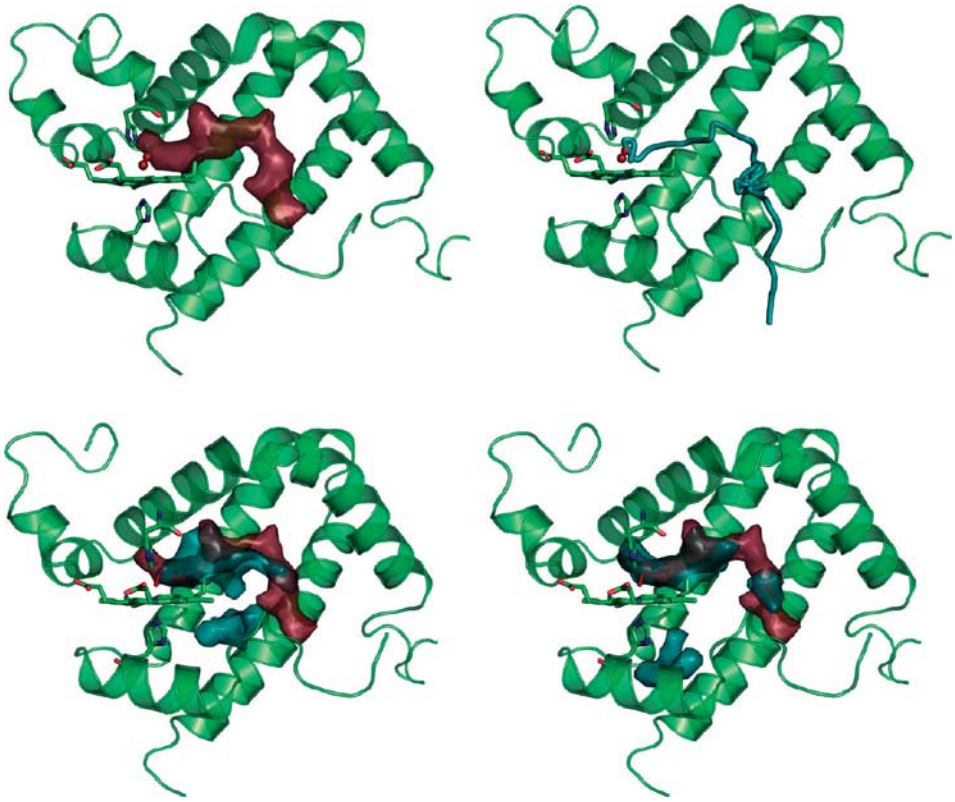


Figure 10. Top: Representation of the cavities found from FPOCKET computations for the oxy state of wt AHb1 (left) and pathway found for ligand migration from MD-GRID computations (right). Bottom: Superposition of the cavities found for the oxy forms of AHb1 (in dark red) and its PheB10 → Leu (left) and HisE7 → Leu (right) mutants. For the sake of clarity, helix E is not shown.

For the oxy form of the HisE7 → Leu mutant, a small cavity lined by residues Ala81, Trp141, Gly142, and Ala144 was formed mimicking the secondary cavity observed in the wt protein (Figure 10). Again, the formation of this cavity was related to fluctuations in the side chain of Tyr145, which was transiently hydrogen-bonded to Phe114 (distances ranging from 2.4 to 6.6 Å; see Figure S8 in the Supporting Information). Furthermore, a second cavity (lined by residues Leu100, Val109, Phe114, Val149, Ile152, and Met156) was found below the heme, which at some extent overlaps with that found in the PheB10 → Leu mutant. Again, this pocket can act as a temporary site for diatomic ligands, thus explaining the larger geminate rebinding observed relative to the wt protein.

Functional Implications. All of the experimental and computational data discussed in the preceding sections reveal a delicate balance of structural effects related to the presence of HisE7 and PheB10 residues in the distal cavity of wt AHb1.

As expected, mutation of distal HisE7 to leucine leads to the formation of a protein that exists as a pure pentacoordinated species. In contrast, a mixture of penta- and hexacoordinated forms is observed for the PheB10 → Leu mutant, thus reflecting the crucial role played by this residue in mediating the equilibrium between 5cHS and 6cLS species. However, whereas the wt protein predominates in the hexacoordinated form, the reverse is found for the PheB10 → Leu mutant. Thus, the equilibrium constant between 5cHS and 6cLS species changes from 1.12 for the wt AHb1 to around 0.59 for the PheB10 → L mutant. These findings point out a subtle but significant role played by PheB10 in modulating the equilibrium between 5cHS and 6cLS species in wt AHb1, as the replacement of this residue by leucine weakens the tendency of HisE7 to bind the heme Fe. In turn, the interplay between HisE7 and PheB10 likely explains the much higher apparent bimolecular rebinding rate in the mutants relative to the wt AHb1. Clearly, this finding is not unexpected for the HisE7 → Leu mutant, as it is found to populate the pentacoordinated species. However, the increased rate of the bimolecular phase in the PheB10 → Leu mutant can be explained by the enhanced preference for the pentacoordinated form, suggesting that ligation and detachment of the distal His occurs with higher rates/lower yield for this mutant.

The subtle interplay between HisE7 and PheB10 is also revealed upon inspection of the spectroscopic data collected for CO and CN^- complexes of AHb1 and its mutated proteins. The differences in the $\nu_{(FeC)}$ and $\nu_{(CO)}$ frequencies observed for the wt protein and its HisE7 → Leu mutant provide clear evidence for the direct involvement of HisE7 in assisting the heme-bound CO through the formation of a hydrogen bond in wt AHb1. On the other hand, the role played by PheB10 is less apparent. Thus, spectral data for CO complexes suggest the involvement of two conformers characterized by different trends in the $\nu_{(FeC)}$ and $\nu_{(CO)}$ frequencies: while in one case $\nu_{(FeC)}$ and $\nu_{(CO)}$ frequencies at 493 and 1965 cm^{-1} are indicative of reduced polar interactions between CO and the surrounding distal residues, in the other case, there is a shift of the $\nu_{(FeC)}$ frequency to 519 cm^{-1}, which unexpectedly is not accompanied by a concomitant upshift of the $\nu_{(CO)}$ frequency. These findings suggest the occurrence of a

subtle rearrangement of the distal His relative to the heme arising from the reduced steric hindrance of leucine in the PheB10 → Leu mutant, leading to a weakening of the electrostatic interaction with the heme-bound ligand. In turn, it can be suggested that PheB10, besides affecting the 5c ⇔ 6c equilibrium through interaction with HisE7, could contribute to the stabilization of the heme-bound ligand by indirect assistance to the formation of the hydrogen bond with HisE7.

The enhanced geminate rebinding phase observed in the two mutants compared to the wt protein can be interpreted from the rearrangements observed in the internal cavities from MD simulations. In particular, it is worth noting that simulations carried out for the oxyferrous form of wt AHb1 clearly delineate a tunnel leading from the distal cavity to the bulk solvent. In contrast, Figure 10 reveals the existence of a series of disconnected cavities in the two mutants, which would thus disfavor migration of ligands through the protein matrix to the solvent compared to the wt protein. Rather, the ligand might be easily trapped in those cavities, which would facilitate recombination with the heme. Figure 10 also shows that the nature of those cavities is not identical for the two mutants. In particular, the cavity found below the heme appears to be more isolated in the HisE7 → Leu mutant than the corresponding cavity in the PheB10 → Leu one. If one assumes the temporary occupancy of these sites by the photodissociated ligand, the location of this cavity would likely explain the different temperature sensitivity of the geminate phase in the two mutants. Thus, the closer proximity of the cavities in the PheB10 → Leu mutant would facilitate migration of the ligand to the distal cavity, as suggested by the higher value of k_{-c} observed for the PheB10 → Leu mutation. In contrast, the larger separation of the cavity observed for the HisE7 → Leu mutant implies a smaller value of k_{-c} (Table 2).

Finally, it is worth discussing the preceding findings in the context of the putative NO detoxification role recently suggested for AHb1, which would act as a NO scavenger. The results suggest that the distal residues HisE7 and PheB10 might have a direct implication in mediating the balance between penta- and hexacoordinated species but also in assisting ligand binding to the heme. NO levels increase in response to hypoxia in plants.[9] This increase appears to be modulated by levels of AHb1, which also increase in abundance in response to hypoxia.[5,6] Keeping in mind the extremely high affinity of AHb1 for oxygen (K_d ∼2−10 nM) and the expression of AHb1 induced by low levels of oxygen,[5,6] which then stimulates production of NO, it can be hypothesized that the interplay of interactions between distal residues HisE7 and PheB10 acts in a synergic way in order to ensure functional activation of AHb1 only under those conditions. Even if the protein tends to rest *under latency* due to the preference toward the hexacoordinated species, the high O_2 affinity facilitates detachment of the distal His from the sixth coordination position of the heme and displacement toward the pentacoordinated species even under hypoxic conditions, thus rendering the *active* oxyferrous species able to scavenge NO. Binding of O_2 would facilitate the opening of the tunnel, leading to the solvent through the secondary cavity (Figure 10), thus allowing the migration of NO as an incoming ligand from the bulk solvent to the heme-bound O_2 in the distal cavity, or involving the formation of a Cys−NO adduct, as suggested by Perazzolli and co-workers.[12]

According to this mechanism, it might be suggested that the subtle interplay between HisE7 and PheB10 has been designed to ensure the access of NO after oxygenation of wt AHb1, thus enabling the protein to accomplish the NO scavenging role. Interestingly, a conceptually related functional mechanism has been recently proposed for the NO dioxygenase activity of truncated hemoglobin N (trHbN) from *M. tuberculosis*. This protein has been hypothesized to facilitate access of NO to the heme cavity in the oxygenated state, as binding of O_2 to the Fe would trigger the opening of the gate (played by the PheE15 residue) that modulates migration of diatomic ligands through a hydrophobic tunnel.[47,48] Therefore, in spite of the notable structural dissimilarities observed between AHb1 and trHbN, it seems that their structures have evolved in order to facilitate migration of NO to the heme cavity only when the protein is ready to accomplish the dioxygenase role. Overall, the preservation of this functional mechanism in stress-induced hemoglobins could reflect a strategy designed by evolution to enhance the efficiency of the NO scavenging activity.

As a final remark, present results trigger challenging questions about the potential role played by distal residues HisE7 and PheB10 in mediating the balance between penta- and hexacoordinated species. The sequence identity between AHb1 and other hexacoordinated proteins, such as neuroglobin and cytoglobin, is very low (around 19 and 24%, respectively), which reflects the evolutionary divergence between these proteins. It has been suggested that the intrinsic conformational flexibility of the CD−D region in globins capable of heme endogenous hexacoordination can be related to the ability of exploiting the 5c ⇔ 6c equilibrium for the control of O_2 diffusion to/from the heme.[49,50] In this context, future studies addressing specifically the differences in the dynamical behavior of wt AHb1 and its PheB10 → Leu mutant might be valuable to shed light into the structural basis and functional roles associated with heme hexacoordination.

Conclusions

The analysis of experimental and computational results reveals the critical influence played by HisE7 and PheB10 residues in the distal cavity on the structure and function of AHb1. Thus, the delicate interplay between HisE7 and PheB10 residues appears to be crucial to modulate the balance between penta- and hexacoordinated forms of the wt protein. Moreover, they would also affect the binding of the exogenous ligand (e.g., O_2), which in turn would facilitate, upon binding, the formation of a passage through the protein matrix leading from the distal cavity to the bulk solvent, thus opening an easier access to a second reactant (e.g., NO). Overall, these findings are consistent with the putative NO detoxification role proposed for AHb1. The synergism between HisE7 and PheB10 would thus be conceived as a subtle mechanism to modulate activation of the protein only when the level of oxygen has surpassed a threshold value, making the oxyferrous form able to perform NO scavenging.

Acknowledgment. The authors acknowledge MIUR (PRIN2004), the Spanish Ministerio de Innovación y Ciencia (SAF2008-05595), the Generalitat de Catalunya (SGR2009294 and XRQTC), and the EU (FP7 Framework Program; project NOstress) for financial support. This work has been performed under the Project HPC-EUROPA++ project (project number: 1076), with the support of the European Community - Research Infrastructure Action of the FP7 "Coordination and support action" Programme. The Biostructures and Biosystems National Institute is acknowledged for the fellowship to F.S. The Barcelona Supercomputer Center is kindly acknowledged for the facilities provided in the *MareNostrum* supercomputer.

Supporting Information Available: Determination of the equilibrium constant for hexacoordination of PheB10 → Leu AHb1. Resonance Raman spectra for CN^- adducts with isotope substitution, time evolution of potential energy and rmsd along the trajectories sampled in molecular dynamics simulations, and of certain structural parameters for specific interactions between residues. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Arredondo-Peter, R.; Hargrove, M. S.; Sarath, C.; Moran, J. F.; Lohrman, J.; Olson, J. S.; Klucas, R. V. *Plant Physiol.* **1997**, *115*, 1259.
(2) Duff, S. M. G.; Wittenberg, J. B.; Hill, R. D. *J. Biol. Chem.* **1997**, *272*, 16746.
(3) Smagghe, B. J.; Sarath, G.; Ross, E.; Hilbert, J. L.; Hargrove, M. S. *Biochemistry* **2006**, *45*, 561.
(4) Wang, Y. H.; Kochian, L. V.; Doyle, J. J.; Garvin, D. F. *Plant, Cell Environ.* **2003**, *26*, 673.
(5) Trevaskis, B.; Watts, R. A.; Andersson, C. R.; Llewellyn, D. J.; Hargrove, M. S.; Olson, J. S.; Dennis, E. S.; Peacock, W. J. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 12230.
(6) Watts, R. A.; Hunt, P. W.; Hvitved, A. N.; Hargrove, M. S.; Peacock, W. J.; Dennis, E. S. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *98*, 10119.
(7) Arredondo-Peter, R.; Hargrove, M. S.; Moran, J. F.; Sarath, G.; Klucas, R. V. *Plant Physiol.* **1998**, *118*, 1121.
(8) Hill, R. D. *Can. J. Bot.* **1998**, *76*, 707.
(9) Dordas, C.; Rivoal, J.; Hill, R. D. *Ann. Bot.* **2003**, *91*, 173.
(10) Wang, R.; Guegler, K.; LaBrie, S. T.; Crawford, N. M. *Plant Cell* **2000**, *12*, 1491.
(11) Nie, X.; Hill, R. D. *Plant Physiol.* **1997**, *114*, 835.
(12) Perazzolli, M.; Dominici, P.; Puertas, M. C. R.; Zago, E.; Zeier, J.; Sonoda, M.; Lamb, C.; Delledonne, M. *Plant Cell* **2004**, *16*, 2785.
(13) Perazzolli, M.; Romero-Puertas, M. C.; Delledonne, M. *J. Exp. Bot.* **2006**, *57*, 479.
(14) Belenghi, B.; Romero-Puertas, M. C.; Vercammen, D.; Brackenier, A.; Inze, D.; Delledonne, M.; Breusegem, F. V. *J. Biol. Chem.* **2007**, *282*, 1352.
(15) Bruno, S.; Faggiano, S.; Spyrakis, F.; Mozzarelli, A.; Abbruzzetti, S.; Grandi, E.; Viappiani, C.; Feis, A.; Mackowiak, S.; Smulevich, G.; Cacciatori, E.; Dominici, P. *J. Am. Chem. Soc.* **2007**, *129*, 2880.
(16) Smagghe, B. J.; Kundu, S.; Hoy, J. A.; Halder, P.; Weiland, T. R.; Savage, A.; Venugopal, A.; Goodman, M.; Premer, S.; Hargrove, M. S. *Biochemistry* **2006**, *45*, 9735.
(17) Goodman, M. D.; Hargrove, M. S. *J. Biol. Chem.* **2001**, *276*, 6834.
(18) Hargrove, M. S.; Brucker, E. A.; Stec, B.; Sarath, G.; Arredondo-Peter, R.; Klucas, R. V.; Olson, J. S.; Phillips, G. N., Jr. *Structure* **2000**, *8*, 1005.
(19) Abbruzzetti, S.; Bruno, S.; Faggiano, S.; Grandi, E.; Mozzarelli, A.; Viappiani, C. *Photochem. Photobiol. Sci.* **2006**, *5*, 1109.
(20) Steinbach, P. J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1476.
(21) Steinbach, P. J.; Ionescu, R.; Matthews, C. R. *Biophys. J.* **2002**, *82*, 2244.
(22) Abbruzzetti, S.; Bruno, S.; Faggiano, S.; Ronda, L.; Grandi, E.; Mozzarelli, A.; Viappiani, C. *Methods Enzymol.* **2008**, *437*, 329.
(23) Bruno, S.; Faggiano, S.; Spyrakis, F.; Mozzarelli, A.; Cacciatori, E.; Dominici, P.; Grandi, E.; Abbruzzetti, S.; Viappiani, C. *Gene* **2007**, *398*, 224.
(24) Hoy, J. A.; Robinson, H.; Trent, J. T.; Kakar, S.; Smagge, B. J.; Hargrove, M. S. *J. Mol. Biol.* **2007**, *371*, 168.
(25) Laskowski, R. A.; MacArthur, M. W.; Mos, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283.
(26) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
(27) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; rozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, CA, 2006.
(28) Martí, M. A.; Crespo, A.; Capece, L.; Boechi, L.; Bikiel, D. E.; Scherlis, D. A.; Estrín, D. A. *J. Inorg. Biochem.* **2006**, *100*, 761.
(29) Bidon-Chanal, A.; Martí, M. A.; Crespo, A.; Milani, M.; Orozco, M.; Bolognesi, M.; Luque, F. J.; Estrin, D. A. *Proteins* **2006**, *64*, 457.
(30) LeGuilloux, V.; Schmidtke, P.; Tuffery, P. *BMC Bioinf.* **2009**, *10*, 168.
(31) Carrillo, O.; Orozco, M. *Proteins* **2008**, *70*, 892.
(32) Dewilde, S.; Kiger, L.; Burmester, T.; Hankeln, T.; Baudin-Creuza, V.; Aerts, T.; Marden, M. C.; Caubergs, R.; Moens, L. *J. Biol. Chem.* **2001**, *276*, 38949.
(33) Aono, S.; Nakajima, H. *Coord. Chem. Rev.* **1999**, 190.
(34) Samuni, U.; Navati, M. S.; Juszak, L. J.; Dantsker, D.; Yang, M.; Friedman, J. M. *J. Phys. Chem. B* **2000**, *104*, 10802.
(35) Spiro, T. G.; Wasbotten, I. H. *J. Inorg. Biochem.* **2005**, *99*, 34.
(36) Phillips, G. N. J.; Teodoro, M. L.; Smith, B.; Olson, J. S. *J. Phys. Chem. B* **1999**, *103*, 8817.
(37) Li, T.; Quillin, M. L.; Phillips, G. N. J.; Olson, J. S. *Biochemistry* **1994**, *33*, 1433.
(38) Feis, A.; Santoni, E.; Neri, F.; Ciaccio, C.; De Sanctis, G.; Coletta, M. *Biochemistry* **2002**, *41*, 13264.
(39) Bohórquez, H. J.; Obregón, M.; Cárdenas, C.; Llanos, E.; Suárez, C.; Villaveces, J. L.; Patarroyo, M. E. *J. Phys. Chem. A* **2003**, *107*, 10090.
(40) Spiro, T. G.; Ibrahim, M.; Wasbotten, I. H. In *The Smallest Biomolecules*; Ghosh, A., Ed.; Elsevier: Amsterdam, The Netherlands, 2008; p 96.
(41) Spiro, T. G.; Wasbotten, I. H. *J. Raman Spectrosc.* **2005**, *34*, 725.
(42) Kushkuley, B.; Stavrov, S. S. *Biochim. Biophys. Acta* **1997**, *1341*, 238.
(43) Deng, T. J.; Macdonald, I. D. G.; Simianu, M. C.; Sykora, M.; Kincaid, J. R.; Sligar, S. G. *J. Am. Chem. Soc.* **2001**, *123*, 269.
(44) Das, T. K.; Couture, M.; Guertin, M.; Rousseau, D. L. *J. Phys. Chem. B* **2000**, *104*, 10750.
(45) Tanaka, T.; Yu, N.-T.; Chang, C. K. *Biophys. J.* **1987**, *52*, 801.
(46) Merchant, K. A.; Noid, W. G.; Akiyama, R.; Finkelstein, I. J.; Goun, A.; McClain, B. L.; Loring, R. F.; Fayer, M. D. *J. Am. Chem. Soc.* **2003**, *125*, 13804.
(47) Crespo, A.; Marti, M. A.; Kalko, S. G.; Morreale, A.; Orozco, M.; Gelpi, J. L.; Luque, F. J.; Estrin, D. *J. Am. Chem. Soc.* **2005**, *127*, 4433.
(48) Bidon-Chanal, A.; Marti, M. A.; Estrin, D. A.; Luque, F. J. *J. Am. Chem. Soc.* **2007**, *129*, 6782.
(49) Pesce, A.; Dewilde, S.; Nardini, M.; Moens, L.; Ascenzi, P.; Hankeln, T.; Bolognesi, T. B. *Structure* **2003**, *11*, 1087.
(50) Capece, L.; Marti, M. A.; Bidon-Chanal, A.; Nadra, A.; Luque, F. J.; Estrin, D. A. *Proteins* **2009**, *75*, 885.

JP9074477