Structure and evolution of protein allosteric sites

by
Alejandro Panjkovich

Thesis submitted to
Universitat Autònoma de Barcelona
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Director - Prof. Xavier Daura
Tesi Doctoral UAB/ANY 2013
Ph.D. Program - Protein Structure and Function
Institut de Biotecnologia i Biomedicina

caminante, no hay camino,
se hace camino al andar

Antonio Machado, 1912

Acknowledgements

First of all I would like to thank my supervisor and mentor Prof. Xavier Daura for his consistent support and trust in my work throughout these years. Xavi, I deeply appreciate the freedom you gave me to develop this project while you were still carefully aware of the small details. Working under your supervision has been a rich and fulfilling experience.

Of course, thanks go as well to current and past members of our institute, especially Rita Rocha, Pau Marc Muñoz, Oscar Conchillo, Dr. Martín Indarte, Dr. Mario Ferrer, Prof. Isidre Gibert, Dr. Roman Affentranger and Dr. Juan Cedano for their technical and sometimes philosophical assistance. Help from the administrative staff was also significant, I would like to thank in particular Eva, Alicia and Miguel who where always ready to help me in sorting out unexpected bureaucratic affairs.

I would also like to thank Dr. Mallur Srivatasan Madhusudhan and his group (especially Kuan Pern Tan, Dr. Minh Nguyen and Binh Nguyen), and also Dr. Gloria Fuentes, Cassio Fernandes, Youssef Zaki, Thijs Kooi, Rama Iyer, Christine Low and many others at the Bioinformatics Institute BII - A*STAR in Singapore for the many interesting discussions and support during my stage over there.

Special thanks go to my previous long-term supervisors Prof. Francisco Melo and Prof. Patrick Aloy, together with friends and colleagues in research Prof. Marc Martí-Renom, Dr. Andreas Zanzoni, Dr. Amelie Stein, Prof. Andrej Sali, Prof. Olivier Michielin, Dr. Aurélien Grosdidier, Dr. Madan Babu, Dr. Damien Devos, Dr. Roland Pache and Dr. Roberto Mosca who among others have played a crucial role in shaping my enthusiasm for science.

I thank as well those beyond the scientific realm but very close to my heart, thanks for every moment shared during this adventure, you know who you are: Lara, Claudia, Jorge, Mary, Manfred, Kiki, Iani, Feli, Eva and the whole family spread over Chile, Argentina, Switzerland and Germany including those who have left, and those who are on their way.

For the many lessons and amazing experiences during these years I am deeply grateful to my dear friends Charlie, Max, Elisa, Gloria, Susana, Rita, Pau, Carlos, Paula, Fran, Isaac, Iván, Manu, Marcela, Consuelo, Mariona, Cata, Sofía, Nicolás, Lucas, Ignacia, Shaf, Sleena, Aurora, Valentina, Amandine, Àgueda, Gemma, Laura, Patri, Clara, Anna, Elena and Pablo among many others.

Last but not least, I acknowledge the financial support I received in the form of a Ph.D. Scholarship 'Formación de Profesorado Universitario' (FPU) reference: AP2008-03046, granted by the Spanish Ministry of Education.

Abstract

This thesis studies protein allosteric sites from a structural and evolutionary perspective. Allostery is a fundamental aspect of life at the molecular level, the most common and powerful mechanism of protein activity regulation: through binding of a ligand to a site which is not the active site. This phenomenon was first described more than 50 years ago and it still captures the attention of researchers, while fully understanding its mechanisms remains a grand scientific challenge. Furthermore, allosteric sites have been increasingly calling the attention of medicinal chemists and pharmaceutical companies, given their potential for the development of novel therapeutics.

The thesis is presented as a 'compendium of published articles'. The first article was published at the beginning of 2010 describing the first stage of the thesis, a series of large-scale computational analyses to characterize putative small-molecule binding sites by integrating publicly available information on protein sequences, structures and active sites for more than a thousand protein families. By identifying common pockets across different structures of the same protein family a method was developed to measure the pocket's structural conservation. Characterization of putative ligand-binding sites followed using different measures such as sequence conservation, structural flexibility, electrostatic potential and structural conservation. The most relevant finding was the unexpected lack of correlation between the two conservation measures, of sequence and structure, for many of the predicted cavities. This general finding was also observed in specific cases of allosteric proteins, where the active site was conserved both in terms of structure and sequence but the allosteric site was conserved only from the structural perspective and did not show conservation at the sequence level.

The second article was published at the end of 2012, it explores the relationship between protein flexibility and allosteric effects defining a computational methodology to predict the presence and location of allosteric sites on protein structures. Besides the dynamical aspects assessed through normal mode analysis, the method also incorporates the structural conservation measure defined in the first article. The predictive approach was benchmarked against a large data set of allosteric proteins of known structure obtaining 65% positive predictive value.

After the second publication, the method has been implemented in the form of a freely available web-server aimed to support the work of researchers in the field of allosterism, both to improve the understanding of this fundamental form of protein function regulation and to serve applied purposes in the area of drug design and discovery.

Resumen

La presente tesis estudia los sitios alostéricos desde una perspectiva estructural y evolutiva. La regulación alostérica es un aspecto fundamental de la vida a nivel molecular, ya que es el mecanismo más potente y frecuente en la regulación de la actividad proteica: mediante la unión de un ligando a un sitio que no es el sitio activo. Este fenómeno fue descrito por primera vez hace más de 50 años y desde entonces no ha dejado de captar la atención de la comunidad científica, llegando incluso a ser calificado como 'el segundo secreto de la vida', después del código genético. Sin embargo, la comprensión cabal de los mecanismos involucrados continúa siendo un gran desafío científico. Actualmente, los sitios alostéricos han despertado un creciente interés por parte de expertos en química medicinal y compañías farmacéuticas, dado su potencial para el desarrollo de nuevos fármacos.

La tesis se presenta como un 'compendio de publicaciones'. El primer artículo fue publicado a principios del año 2010 y describe la primera etapa del proyecto, una serie de análisis computacionales a gran escala para caracterizar sitios de unión a ligando integrando información referente a secuencia, sitios activos y estructura para más de mil familias proteicas. Mediante la identificación de sitios de unión comunes en distintas estructuras de la misma familia proteica, se desarrolló un método para medir la conservación estructural de dichos sitios. Esta metodología permitió realizar una caracterización de sitios de unión considerando distintos aspectos, como la conservación evolutiva a nivel de secuencia, flexibilidad estructural, potencial electrostático y conservación estructural. El descubrimiento más significativo fue la inesperada falta de correlación entre las medidas de conservación de secuencia y estructura para muchos de los sitios de unión predichos. Este hallazgo es válido también para casos específicos de proteínas alostéricas, donde el sitio activo está conservado tanto a nivel de secuencia como de estructura, pero el sitio alostérico sólo presenta conservación a nivel estructural y no de secuencia.

El segundo artículo fue publicado a fines del año 2012 y explora la relación entre la flexibilidad proteica y la regulación alostérica, definiendo una metodología computacional para la predicción de sitios alostéricos en estructuras proteicas. Más allá de los aspectos dinámicos que fueron estudiados mediante el análisis de modos normales, el método también incorpora la medida de conservación estructural desarrollada en el primer artículo. El sistema predictivo fue puesto a prueba utilizando un extenso conjunto de proteínas alostéricas de estructura conocida, obteniendo un valor predictivo positivo de 65%.

Después de la segunda publicación, el método se ha implementado como servidor web para brindar apoyo a la investigación de la regulación alostérica, tanto para extender el conocimiento de esta forma fundamental de regulación de la actividad proteica, como para ayudar en la aplicación de dichos conocimientos al desarrollo de nuevos fármacos con objetivos terapéuticos.

Contents

Chapter 1

Objectives

The main objective of this thesis was to study allostery from a structural and evolutionary perspective. This serves a dual-purpose, on one hand we can extend the academic understanding of allosteric phenomena and, on the other, such understanding may pave the way for the discovery and further development of novel therapeutics. The scientific objectives of this endeavour were to:

- Characterize putative ligand-binding sites in terms of structural and evolutionary aspects.
- Explore the relationship between allosteric ligand binding and protein flexibility.
- Design and implement a computational method to predict allosteric sites.

The first objective was addressed in the first part of the project and is elaborated in Article I, where in order to study a particular ligand-binding site or pocket across different structures of the same protein family it was necessary to develop a special measure of structural conservation.

The second and third objectives were approached in the second part of the project and are elaborated in Article II. The methodology was developed to support the first step of a drug-design pipeline by selecting targets for further experimental characterization. In that context, it was important to implement a method that was fast in execution, meaning that large sets of protein structures could be analyzed in a reasonable amount of time.

Chapter 2

Introduction

Life can be regarded as a highly complex and organized molecular phenomenon. The prime image of self-organized life, which also illustrates the relationship between function and structure at the molecular level, is DNA (deoxyribonucleic acid). The elucidation of its self-replicating structural features was crucial for understanding the molecular basis of inheritance and evolution 60 years ago (Watson and Crick, 1953). However, even if genes as units of biological information are coded into DNA, it is their translation into protein molecules what transforms abstract information into chemical action. One may say that gene is substantive and protein is verb. Thus, proteins can be considered the functional units of life as they carry out and coordinate almost all biological processes. In fact, a myriad of chemical reactions and physical rearrangements take place within a single living organism every day. Not only is their number large, these processes are highly intertwined. As such, and contributing to their overall intricacy, physicochemical processes related to life are characterized by implicit regulation, which is a fundamental part of their self-organizing nature.

2.1 Allostery

In this thesis, I have studied the most common, direct and powerful means of protein function regulation, allostery. A protein is under allosteric regulation when an effector molecule alters its function by binding at a site which is not the active site. Usually key proteins in signalling-cascades and/or metabolic pathways are regulated allosterically, allowing immediate responses to changes in their chemical environment.

2.1.1 Historical background

The first findings related to the regulation of protein activity can be dated back to the discovery of the sigmoidal binding curve of Oxygen to hemoglobin by Christian Bohr in 1904 (Christian Bohr and Krogh, 1904) and the formulation of the Hill equation to describe it (Hill, 1910). The key aspect is that the affinity of hemoglobin for O_2 changes according to its concentration and other environmental aspects such as pH and $[CO_2]$ (Perutz, 1970). Decades later, Cori and coworkes described the regulatory effect of adenosine monophosphate on glycogen phosphorylation (Cori GT and CF., 1938). During the years that followed researchers were busy drawing metabolic pathways without much concern on how the involved reactions were kept under control by the cell, until the discovery of feedback inhibition (Yates and Pardee, 1956; Kresge *et al.*, 2005). Until then, inhibitors were generally thought to bind the enzyme's active site and compete with the substrate. However, these new discoveries included inhibitors that could differ from substrates in terms of shape, size and charge. Scientist were puzzled, if the inhibitor was chemically distinct from the substrate, how could it compete for binding to the active site?

The answer came through the concept of allostery developed by Jacques Monod *et al.* half a century ago, as they described how regulation could take place at a binding-site distinct from the active site (Monod and Jacob, 1961; Monod *et al.*, 1963; Monod *et al.*, 1965).

2.1.2 Etymology and definition

Nussinov and coworkers state that "allostery is regulation at a distance" in a recent review (Nussinov and Tsai, 2013). They go on to explain that allostery "is a universal phenomenon whereby a perturbation by an effector at one site of the molecule leads to a functional change at another through alteration of shape and/or dynamics."

This definition matches the explanation one commonly finds for the word's etymology, which comes from the Greek *allos*, "other", and *stereos*, "solid / object", in reference to the fact that the regulatory site of an allosteric protein is located in a different (non-overlapping) position respect to its active site (Helmstaedt *et al.*, 2001; Changeux, 2011).

However, a slightly different definition can be found on other sources, for example a classic textbook on molecular biology (Alberts *et al.*, 2002):

> "One feature of feedback inhibition was initially puzzling to those who discovered it: the regulatory molecule often has a shape totally different from the shape of the substrate of the enzyme. This is why this form of regulation is termed allostery (from the Greek words allos, meaning 'other,' and stereos, meaning 'solid' or 'three-dimensional'). As more was learned about feedback inhibition, it was recognized that many enzymes must have at least two different binding sites on their surface,

> an active site that recognizes the substrates, and a regulatory site that recognizes a regulatory molecule. These two sites must somehow communicate in a way that allows the catalytic events at the active site to be influenced by the binding of the regulatory molecule at its separate site on the protein's surface."

So, does the *steric* component of *allosteric* refer to the shape of the effector molecule or to its binding site on the protein? Of course these are equivalent to some degree, since a differently shaped effector molecule would in principle bind a different protein pocket. However, it is interesting to see how these slightly diverse definitions evolved by looking at the original publications. This rather historical exercise will help in understanding the protein-centered focus of the work presented in this thesis. Jacques Monod and François Jacob coined the *allosteric* term more than 50 years ago when they wrote the summary article of the 26th Cold Spring Harbor Symposium on Quantitative Biology titled "cellular regulatory mechanisms" (Monod and Jacob, 1961). We can look at their exact wording to understand the evolution of the terminology in its context:

> "As the reports have shown, endproduct inhibition is extremely widespread in bacteria, insuring immediate and sensitive control of the rate of metabolite biosynthesis in most, if not all, pathways. From the point of view of mechanisms, the most remarkable feature of the Novick-Szilard-Umbarger effect is that the inhibitor *is not a steric analogue of the substrate*. We propose therefore to designate this mechanism as 'allosteric inhibition.' Since it is well known that competitive behavior toward an enzyme is, as a rule, restricted to steric analogues, it might be argued that an enzyme's concept of steric analogy need not be the same as ours, and that proteins may see analogies where we cannot discern any."

The concept was developed and more precisely defined two years later in a following article by Monod *et al.* (Monod *et al.*, 1963). Most notably, the term's focus shifts from the effector molecule towards the protein molecule as we read:

> "The specificity of any allosteric effect and its actual manifestation is therefore considered to result exclusively from the specific construction of the protein molecule itself, allowing it to undergo a particular, discrete, reversible conformational alteration, triggered by the binding of the allosteric effector."

This idea is then further emphasized in the same publication:

> "Finally, the coexistence in *E. coli* of two different aspartokinases catalysing identical reactions, respectively inhibited by threonine and by lysine (Stadtman *et al.*,

> 1961), offers a striking illustration of the fact that the nature and structure of the inhibitor is, in a sense, irrelevant to the interpretation of the effect. Clearly, such an interpretation must be sought exclusively in the functional structure of the regulatory protein itself."

The crucial observation for rationalizing allosteric effects was the fact that proteins could be *desensitized* to allosteric regulation through mutations while keeping their catalytic function intact. Beyond mutations simply altering the regulatory binding site, the authors observed that regulation mechanisms were more sensitive to mutations than the catalytic function, meaning that allosteric transitions involved not only the regulatory site but also relied on residues spread over the protein structure.

2.1.3 Classic models

Monod, Wyman and Changeux published the first model of allosteric regulation in 1965, it is normally referred to as the MWC model (following the authors initials) and also known as the *concerted* model (Monod *et al.*, 1965). A year later an alternative model was presented by Koshland, Némethy and Filmer, dubbed KNF or *sequential* model (Koshland *et al.*, 1966).

Both phenomenological models were designed to explain the now classic example of allosteric regulation and cooperative binding where a ligand bound to one subunit alters ligand affinity on the rest of the subunits composing the protein assembly as observed in the case of hemoglobin (Pauling, 1935). The models assume that each subunit has at least one ligand binding site and can adopt either a relaxed (R) or tense (T) state, with a large difference in ligand affinity between states.

The MWC model further postulates that subunits in the protein complex are identical and change conformation in a concerted manner, defining the conservation of symmetry as a fundamental requirement. Thus, according to this model, all subunits in each oligomer are either in the R or T state. In solution, the different conformational states of the allosteric protein oligomer coexist in equilibrium and are explored independently of the presence of the allosteric ligand. The cooperativity effect occurs because, upon ligand binding to a single subunit, the rest of subunits and unbound binding sites on the oligomer are stabilized in the higher affinity R state, effectively shifting the equilibrium of conformational states and, in turn, overall ligand affinity. The authors of this model also distinguish between *homotropic* regulation where the effector ligand is identical to the substrate and *heterotropic* interactions where the effector is a different molecule. The MWC model has been applied on multiple cases ranging from hemoglobin to membrane receptors (Changeux and Edelstein, 2005).

While the MWC model is close to what today is referred to as *conformational selection*,

because the population of conformations explored by the protein is shifted upon ligand binding, the KNF model is based on the concept of *induced-fit*, where the ligand induces a conformational change upon binding. Contrary to the MWC model, the KNF model considers that subunits may undergo conformational change independently of the other subunits upon ligand binding. However, ligand binding to one of the subunits in the oligomer induces a conformational change that facilitates neighboring subunits to switch conformation to the high-affinity state explaining in this way the cooperativity effect.

The point of apparent disagreement between both models is whether the conformational change occurs prior or after ligand binding, which was conciliated through a general physical framework by Weber (Weber, 1972) and further developments that are described in the next section. For an insightful analysis and comparison of these classic models I refer the reader to the review by Cui and Karplus (Cui and Karplus, 2008).

2.1.4 Modern perspective

Years after the classic models were published and applied to describe multiple systems (Changeux, 2012), allosteric properties were discovered in monomeric proteins leading to the realization that quaternary structure was not a fundamental for allostery (Cardenas *et al.*, 1978; Kamata *et al.*, 2004; Ascenzi *et al.*, 2005). These findings, combined with more developed notions about protein dynamics (McCammon *et al.*, 1977; Frauenfelder *et al.*, 1991) gave rise to a *new view* on allosteric regulation (Kern and Zuiderweg, 2003; Cui and Karplus, 2008; Hilser, 2010). Briefly, the MWC model treats protein subunits as rigid-bodies, while the present understanding considers proteins as dynamic entities which explore distinct conformations in solution forming a population of conformers including active, inactive and intermediate states (*i.e.* the protein may explore the *bound* conformation in the absence of the ligand). The *new view* considers that binding of regulatory ligands simply alters this distribution of conformers effectively creating a *population shift*. As indicated by Cui and Karplus (2008) this definition is not necessarily new, given that conformational selection is part of the MWC model, even though the original model restrains conformational rearrangements to the level of quaternary structure. The current perspective is that these limitations were embraced to enable the construction of the mathematical models needed at the time since intuitively, the concepts behind the *new view* were already present in Monod's 1965 article:

> "By their very nature, allosteric effects cannot be interpreted in terms of the classical theories of enzyme action. It must be assumed that these interactions are mediated by some kind of molecular transition (allosteric transition) which is induced or stabilized in the protein when it binds an *allosteric ligand*."

This return to a more fundamental definition is remarkable given that allostery pioneers did not have access to the vast amount of information on protein structures and dynamics, genomic sequences and others which the last decades have brought, but they were still able to foresee what today we considered the main principles behind allosteric mechanisms. Laskowski *et al.* distinguish between allosteric effectors that produce a *population shift* and those that directly change the flexibility of a protein, thus altering its activity (Laskowski *et al.*, 2009). Even though thermodynamic and kinetic factors can be treated separately, they are related given that the amount of distinct conformations a protein explores will depend on its intrinsic flexibility, *i.e* the height of the barriers between these conformations. At present we understand that if a protein's flexibility is affected by the binding of a ligand, this may indeed affect its activity rate as stated by Peracchi and Mozzarelli:

> "In a generalized way, any event that perturbs a protein at an allosteric site generates an allosteric signal leading ultimately to an allosteric transition that affects function(s) at distant site(s). Thus, a key requirement for allostery is protein flexibility."

Moreover, this has led to the idea that all (nonfibrous) proteins may be regulated allosterically (Gunasekaran *et al.*, 2004). Furthermore, allosteric signals may also propagate solely by altering protein dynamics, without a detectable conformational change (Cooper and Dryden, 1984; Tsai *et al.*, 2008). As we see, losing the restrictions imposed by the initial models expands the reach of the allosteric phenomenon while at the same time increases its complexity and diversity, which is the topic of the next section.

2.2 Complexity, diversity and classification of allosteric systems

Human hemoglobin (Hb) was the first allosteric protein to be crystallized and it represents an iconic example of allosteric regulation and cooperativity. Briefly, the binding of O_2 to one subunit (human Hb is a tetramer) increases O_2 affinity on the other monomers. A structural explanation for this phenomenon was delivered in 1970 by Perutz, describing how O_2 binding to one monomer's heme iron resulted in a conformational change on the proximal side (opposite of the heme), which, communicated through the dimer-dimer interface, weakened the T (tense or inactive) state relative to the R (relaxed or active) state (Perutz, 1970). Perutz's interpretation followed the concerted model (MWC) rather than the sequential model from Koshland (KNF), both mentioned above, and his interpretation was considered solid enough at the time. However, in the light of recent experimental data, specialists consider that cooperativity and allostery in Hb follow a more complex mechanism which has not been fully understood yet (Ackers and Holt, 2006).

If the details concerning the mechanism of one of the most studied allosteric systems are not completely clear after decades of work, what is the current understanding of other, less studied systems? Moreover, allosteric systems and mechanisms tend to differ considerably, as Monod noted initially "their physiological diversity is extreme" (Monod *et al.*, 1963). Among this variety, a typical case of *conformational change* upon ligand-binding is exemplified by glucose-induced glucokinase, where the ligand triggers a conformational change which in turn renders the active site functional (Heredia *et al.*, 2006). In other cases protein monomers are not active until assembled into a larger complex and the allosteric ligand may act by stabilizing the protein-protein interfaces, as seen in the case of GTP cyclohydrolase stimulatory complex (Maita *et al.*, 2002). A protein which helps illustrating how varied and complex allosteric mechanisms may be is ribonucleotide reductase. This protein displays two different allosteric sites: one affects the enzyme catalytic rate and the other alters its specificity allowing the enzyme to switch substrates (Reichard, 2010).

Furthermore, allosteric regulation traditionally referred to enzymes where the allosteric effector is a small-molecule ligand. However, the term has been expanded and currently authors may use the term *allosteric* to describe regulation events triggered by binding of another protein molecule, phosphorylation, modification of disulfide bonds or other post-translational modifications on proteins which are not necessarily catalytic enzymes. This extension in the use of the terminology has further increased the complexity of the field.

Recent articles review how protein activity can be affected allosterically through varied mechanisms and classify them following different criteria. For example, Tsai *et al.* classify allosteric effects according to the presence or absence of conformational change at the substrate site and they also distinguish if the effect is enthalpy- or entropy-driven among other criteria (Tsai *et al.*, 2009). With a different perspective and focusing on enzymes, Zorn and collaborators separate cases where the allosteric effector binds directly to the catalytic domain or to a distinct regulatory domain (Zorn and Wells, 2010). Laskowski *et al.* provide a classification scheme which groups cases into different categories according to the nature of the allosteric effect: (1) open/close active site, (2) change active site conformation, (3) change active site electrostatic properties (4) affect protein-protein complex formation, (5) change protein flexibility and (6) *population shift* in ensemble of conformers (Laskowski *et al.*, 2009). Depending on which protein is studied, these classification schemes may overlap while in other cases probably complement each other, more even so if we consider that *population shift* has been postulated as a general explanation for allosteric transitions.

At the moment of writing there seems to be no estimation on how populated each of these mechanisms are, meaning that we do not know which mechanism is most or least common in nature. Actually, the total number of allosteric proteins that exist is not known and has not

been even estimated. The AlloSteric Database (Huang *et al.*, 2011) contains to date 283 allosteric sites at different levels of characterization. However, this number is smaller when redundancy filters are applied. The scarcity of data regarding identification and characterization of allosteric systems is probably related to the different technological breakthroughs which have taken place in the last decades, biasing research to other fields such as DNA sequencing instead of biochemical characterization. The point is that, potentially, many allosteric systems remain to be discovered. Furthermore, researchers have reported the discovery of serendipitous allosteric sites: ligand-binding sites that, despite lacking natural ligands, may become allosteric given the presence of an appropriate 'opportunistic' ligand (Hardy and Wells, 2004; Hardy *et al.*, 2004). In a similar line, Bowman *et al.* describe 'cryptic' allosteric sites which may not be detectable through conventional experiments because of their transient nature (Bowman and Geissler, 2012). These findings support the idea that the vast majority of proteins are in principle prone to allosteric regulation (Gunasekaran *et al.*, 2004). In summary, allosteric systems are complex and varied in nature, remaining far from being fully characterized or understood at present time. A large territory is open for scientific exploration, which is increasingly calling the attention of both academic and applied research as we will see in the next section.

2.3 Pharmacological and therapeutical implications

Unveiling common molecular patterns beneath the variety and complexity of allosteric systems poses a grand scientific challenge which has fascinated researchers since the initial discoveries more than 50 years ago. Besides the long-standing academic interest, during the past decade allosteric systems have been increasingly becoming the focus of applied research by medicinal chemists and pharmaceutical companies. This is fueled mainly by their potential in biomedical applications. In this context, allosteric mechanisms present interesting advantages as targets for the design of novel therapeutics (Nussinov and Tsai, 2013). One of the key factors is that allosteric sites tend to be under lower evolutionary pressure than active sites, facilitating the design of highly specific drugs and reducing the risks of toxicity or side-effects (Moehler *et al.*, 2002; Raddatz *et al.*, 2007; Nussinov *et al.*, 2011). Furthermore, while traditional orthosteric drugs usually bind to the active site and inhibit protein activity, allosteric drugs may not only inhibit but also increase protein activity, enabling novel therapeutic possibilites. An example is the case of Benzodiazepines, which are positive allosteric modulators of GABA receptors used in the treatment of anxiety and sleep disorders without the potentially lethal effects of directly acting GABA receptor agonists (Conn *et al.*, 2009). Traditional drugs may also be complemented by allosteric effectors, as observed in the case of aminoglycoside phosphotransferase where a previously unknown binding site could be exploited to allosterically counteract antibiotic resis-

tance (Kohl *et al.*, 2005). Similarly and as observed in G-protein-coupled receptors (GPCRs), allosteric modulators lacking agonism will not exert their effect in the absence of the orthosteric ligand, meaning that the allosteric drug would be able to maintain the dependence of its effect on endogenous physiological signalling (Conn *et al.*, 2009). In summary, allostery is unveiling a novel territory for drug-design beyond what has been already covered by the classic, active-site oriented drug-development approach (DeDecker, 2000). It is important to mention that sometimes authors describe allosteric events where the allosteric effector is not a small-molecule but a large protein, peptide, DNA, etc. However, as mentioned above, in this work we have chosen to focus on small-molecule binding sites since these are the most common and probably the most appropriate for drug-discovery efforts (Peracchi and Mozzarelli, 2011).

2.4 Previous theoretical work

Understanding allosteric mechanisms from a theoretical perspective has motivated a wide range of approaches, which we will briefly review in this section. The motivation for pursuing such understanding is common across computational or theoretical biology: ideally, if we fully understand a system we can predict its behaviour without (or at least with reduced) experimental work, saving time and valuable resources. Ultimately, possessing such knowledge would enable the design of modification strategies which can be applied on the biological system or living organism with a particular aim, such as biotechnological or therapeutical applications. At the same time, computational approaches help accessing and integrating large quantities of information, a clear advantage when trying to unveil common patterns behind complex phenomena such as protein activity regulation.

Across the field and in this case as well, common data sources for analysis are sequence-level information (corresponding to the residue-sequence or primary structure of proteins) and structural information referring to the atomic three-dimensional arrangement of protein molecules. As we will see below, structural studies may focus on static images of conformational changes or in contrast be more inclined to understand the dynamical component beneath the allosteric transition.

Sequence-level information is usually queried to retrieve evolutionary traits (Schneider, 2000). At its most basic level it allows for positions (amino-acids or residues) in the protein polymer to be distinguished according to their biological relevance in relation to their level of conservation across a protein family. One of the most relevant approaches exploiting evolutionary or sequence-level information in the context of allostery is the work of Rama Ranganathan and collaborators, published in Science in 1999 (Lockless and Ranganathan, 1999). They studied a few allosteric systems from an evolutionary perspective based on the analysis of multiple sequence

alignments (MSA). Briefly, they developed an approach named Statistical Coupling Analysis (SCA) which measures the tendency of certain residues (positions in the MSA) to display correlated substitution patterns. SCA displays significant analogies to other approaches implemented previously to study protein-protein interactions (Lichtarge *et al.*, 1996; Pazos *et al.*, 1997). The work of Ranganathan is particularly interesting because their computational method allowed the identification of pathways of thermodynamically linked residues within the PDZ domain, which would be responsible of signal transmission within the protein structure as confirmed through mutational analysis. Even though their approach succeeds at identifying couplings in the protein folds studied, they explain that

> "it does not reveal the physical mechanism of the energetic coupling. Nevertheless, the arrangement of coupled residues into ordered pathways through the core of the PDZ and POZ protein folds suggests that one mechanism may be simple mechanical deformation of the structure along couple pathways."

Following this thought, it is understandable that many researchers focus their attention on the structural level, which may very well complement evolutionary studies based on sequence analysis. A good example of such complementation can be seen in the work of Olivier Lichtarge and collaborators as they map evolutionary information obtained through sequence analysis onto 3D protein structures allowing the identification and even 'recoding' of positions responsible for the determination of specificity in psychoactive bioamine receptors (Rodriguez *et al.*, 2010).

The main repository of protein structural data is the Protein Data Bank (Berman *et al.*, 2000) and most of its entries consist of 'static' three-dimensional images of protein conformations solved by X-ray crystallography. Clues on the dynamics involved in allosteric transition are scarce in this kind of data, however useful information can be extracted by comparing the conformations of active and inactive proteins. Following this line of thought, Michael Daily and Jeffrey Gray performed an ample study on the conformational changes induced by allosteric signals (Daily and Gray, 2007). They analyzed a set of 51 pairs of known inactive and active allosteric protein structures and concluded that in average 20% of the protein changes local structure during allosteric transition. The number is considerably higher than the 10% average observed for nonallosteric proteins (or so classified) displaying significant motion upon ligand binding. Besides the detailed comparison of structural rearrangements between active and inactive conformations, this work represents the first attempt to compile a structural data set of allosteric proteins. Two years later the same authors published a follow-up article studying both tertiary (residue-scale) and quaternary (domain- and subunit-scale) structural changes on 18 allosteric proteins (Daily and Gray, 2009). In this study they quantified the contribution of tertiary and quaternary rearrangements to the allosteric signal, modelling the geometric change at different levels of each

allosteric protein complex in the form of a 'global communication network'. Their model allowed them to map substrate-effector pathways for 15 out of the 18 proteins analyzed while illustrating how allostery may depend on the connection between small- and large-scale motions. The same year Daily and co-authors published a study aimed at identifying the key residues transmitting the allosteric signal (Demerdash *et al.*, 2009). They tested a set of dynamical and structural, as well as network- and informatic-related features to predict allosteric hot-spots (*i.e.* residues that alter the allosteric response upon mutation) and proceeded to implement these features into a machine-learning approach testing their predictive power on five allosteric proteins. The study shows that the performance of such a predictor is better than the sequence-based SCA method mentioned earlier (Lockless and Ranganathan, 1999) and they propose that one reason may be that evolutionary co-conservation of residues is not necessarily a property specific of allosterically coupled residues. They also comment on the limitiations of other studies mentioned above which do not incorporate dynamical features.

Inherently following the famous quote from Richard Feynman "everything that living things do can be understood in terms of the jigglings and wigglings of atoms" many different research groups agree on the dynamical nature of proteins playing a fundamental role in allosteric phenomena (Freire, 2000; Goodey and Benkovic, 2008; Liu and Nussinov, 2008; Kidd *et al.*, 2009; Rader and Brown, 2011; Nussinov and Tsai, 2013). However, one does not necessarily need to go to the level of atomic detail to study the dynamical features of allosteric systems. In this regard, an extremely simplified approach based solely on dynamical aspects was used to model *Escherichia coli* lac repressor proteins as two plates connected by springs representing the structural domains (Hawkins and McLeish, 2004). In that study, Hawkins and McLeish illustrate how the inducer ligand may affect the intramolecular vibrational entropy thereby altering protein activity. They emphasize the distinction between 'static' conformational changes and changes related to protein dynamics. However, a static structure can be used as a starting point for computational exploration of the protein's conformational space (Freire, 2000). Actually, as Cui and Karplus state "there is a complementarity between structure and dynamics in that the conformational changes that play a functional role in allostery are coded into the structure" (Cui and Karplus, 2008). Nevertheless, this coding of allosteric properties in the protein structure appears to be subtle and fragile when compared for example to a catalytic site. Monod observed this when he described allosteric *desensitation*, explaining that proteins were more prone to loose allosteric properties than catalytic capacity upon mutation (Monod *et al.*, 1963). This may explain the difficulties in predicting or even understanding allosteric properties when compared to catalytic activity, which at present can be predicted systematically with reasonable accuracy (Mistry *et al.*, 2007; Porter *et al.*, 2004). The studies mentioned above have broaden the perspective on how to approach allosteric systems, however at the time none of them provided a readily applicable answer to

basic questions such as "Can protein X be regulated allosterically?" or "Where is the allosteric site located in protein X?" This is due in part to different goals, for example many researchers try to model and understand a particular allosteric system, while others attempt to reach a more general perspective to map the 'pathway' through which the allosteric signal travels within the protein macromolecule from the allosteric to the active site (Lockless and Ranganathan, 1999; Flynn *et al.*, 2003; Ota and Agard, 2005). However, this is questioned by work suggesting that *in vivo* there may be multiple effector sites and multiple signalling paths within a single allosteric protein (del Sol *et al.*, 2009). Furthermore, Hilser considers that precise understanding of the 'signalling pathway' may not be required to exploit allosteric phenomena in an applied scenario, whereas an understanding of the stabilities of native states in the ensemble of conformers may be sufficient (Hilser, 2010). This idea is emphasized by recent findings on intrinsically disordered proteins transmiting an allosteric signal (Hilser, 2013; Ferreon *et al.*, 2013).

One of the first approaches to exploit protein dynamics with a predictive insight was developed by Ming and Wall in 2005 (Ming and Wall, 2005). They studied allosteric effects on protein-ligand pairs by comparing the predicted dynamics of bound and unbound structures. On a second article, they tested their methodology for the prediction of functional ligand-binding sites on a set of 305 protein-ligand complexes, which were not necessarily allosteric (Ming and Wall, 2006). A similar approach has been followed by Mitternacht and Berezovsky that describes a measure called *binding leverage*, which is used to locate biologically relevant binding sites, including allosteric sites. The calculation of *binding leverage* is explained as follows in their article (Mitternacht and Berezovsky, 2011):

> "To find potential binding sites we will employ a minimalistic docking procedure to probe the surface of a protein and generate a list of possible binding sites. For each site we estimate the strain on the ligand-protein contacts under the deformations described by low frequency normal modes. The strain is high when the ligand has many contacts with residues that are moving in opposite directions."

They continue to explain that such a site would have a high binding leverage and that ligands binding such sites present a large potential to affect which states are available to the protein. They applied this analysis on a total of 15 allosteric proteins and observed varied results for specific proteins, concluding that regulatory sites may be identified without previous experimental knowledge on conformational changes. Both the work of Mitternacht and that of Ming and Wall postulate that allosteric sites may be detected by estimating the protein's dynamical perturbation upon ligand-binding to a particular site. This principle is common to our own approach as we will see below.

It is important to mention that during the preparation of this text, a few web-servers have been published, which may be used to predict the location of allosteric sites on protein struc-

tures. The first one published was SPACER (Goncearenco *et al.*, 2013), a web-server based on the *binding leverage* measure and other parameters previously described by Mitternacht *et.al.* (Mitternacht and Berezovsky, 2011) to study allosteric communication across residues in a given protein structure. A slightly different approach but a similar purpose are features of MCPath, a web-server based on Monte Carlo path generation approach and an atomistic potential function to identify residues that have the highest probability of being involved in the allosteric communication pathway (Kaya *et al.*, 2013). A third web-server, in this case aimed directly at predicting the location of allosteric sites, is Allosite (Huang *et al.*, 2013). Allosite has been implemented by the group that built ASD (Huang *et al.*, 2011). The prediction in this case is based on a support vector machine method that has been trained with a non-redundant set of 90 allosteric protein structures.

2.5 Work presented in this thesis

The two articles that represent the fundamental part of this thesis follow the same purpose, the understanding of common patterns underlying allosteric mechanisms. The first article consists of a large-scale characterization of protein pockets and the development of a novel measure of structural conservation (Panjkovich and Daura, 2010), which was then incorporated in the second article as part of a predictive approach. The second article explores how allosteric ligands may perturb protein flexibility upon binding and combines the analysis with structural conservation into a methodolody for predicting the location of allosteric sites on protein structures (Panjkovich and Daura, 2012). The predictive approach was benchmarked on a large set of structurally known allosteric proteins obtaining 65% positive predictive value. Further details are provided in the corresponding articles.

Besides the two published articles that build the fundamental part of this thesis, a shorter manuscript is included as an Annex describing **PARS**, a web-server implementation of our allosteric sites prediction methodology. The web-server is freely accessible at *http://bioinf.uab.cat/pars* making the method readily available for the scientific community. This manuscript was not included in the fundamental body of the thesis because at the time of writing it is under review for publication.

The same applies to the second Annex included, it describes **antibacTR**: another web-based tool designed to prioritize antibacterial drug targets to aid in the design of novel antibacterials. It describes how the methodology developed in this thesis was applied already on the analysis of full bacterial proteomes, involving thousands of protein structures/models, as part of the AntiPathoGN project (Seventh Research Framework Programme of the European Union. ref. HEALTH-F3-2009-223101).

Chapter 3

Results and discussion

Here, I will briefly summarize the results and discussion which are presented in full in the published articles that represent the fundamental body of this thesis (Panjkovich and Daura, 2010; Panjkovich and Daura, 2012).

3.1 Large-scale study of protein pockets and measure of structural conservation

My initial attempts to study the conservation levels of ligand-binding sites across protein families was of limited success, mainly because of the lack of a unified description of protein ligand-binding sites. Characterization values computed for a predicted cavity on one protein structure could not be compared among proteins of the same family if the correspondence to equivalent cavities was not known. Thus, in order to perform a large-scale study, we needed to identify common cavities among different protein structures of the same family. Ideally, we wanted to relate a cavity found in one protein of a given family to the equivalent cavity detected in another protein of the same family. This was the main motivation behind the structural conservation measure implemented in the first article (Panjkovich and Daura, 2010) included in this thesis. Once we could relate pockets across protein families (using structural superimposition (Ortiz *et al.*, 2002) and a specially developed clustering algorithm) we proceeded to characterize protein cavities according to different criteria such as sequence and structural conservation, flexibility and electrostatic potential. After analyzing a filtered total of 22,321 protein structures (involving 4,258 distinct protein families according to Pfam (Finn *et al.*, 2008)) we settled on a total of 1,128 protein families for which we had at least 5 representative structures. We used known active sites as a reference when analyzing the distributions of values (electrostatic potential, flexibility, etc) for the large-scale predicted pockets. The most relevant and unexpected observation was

the lack of correlation between sequence and structural conservation for many of the predicted pockets, this was in hard contrast with active sites where both measures correlated as expected. Close inspection of a few particular cases showed correspondence to allosteric sites.

Panjkovich and Daura *BMC Structural Biology* 2010, **10**:9
http://www.biomedcentral.com/1472-6807/10/9

BMC Structural Biology

RESEARCH ARTICLE Open Access

Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery

Alejandro Panjkovich[1] and Xavier Daura*[1,2]

Abstract

Background: With the classical, active-site oriented drug-development approach reaching its limits, protein ligand-binding sites in general and allosteric sites in particular are increasingly attracting the interest of medicinal chemists in the search for new types of targets and strategies to drug development. Given that allostery represents one of the most common and powerful means to regulate protein function, the traditional drug discovery approach of targeting active sites can be extended by targeting allosteric or regulatory protein pockets that may allow the discovery of not only novel drug-like inhibitors, but activators as well. The wealth of available protein structural data can be exploited to further increase our understanding of allosterism, which in turn may have therapeutic applications. A first step in this direction is to identify and characterize putative effector sites that may be present in already available structural data.

Results: We performed a large-scale study of protein cavities as potential allosteric and functional sites, by integrating publicly available information on protein sequences, structures and active sites for more than a thousand protein families. By identifying common pockets across different structures of the same protein family we developed a method to measure the pocket's structural conservation. The method was first parameterized using known active sites. We characterized the predicted pockets in terms of sequence and structural conservation, backbone flexibility and electrostatic potential. Although these different measures do not tend to correlate, their combination is useful in selecting functional and regulatory sites, as a detailed analysis of a handful of protein families shows. We finally estimated the numbers of potential allosteric or regulatory pockets that may be present in the data set, finding that pockets with putative functional and effector characteristics are widespread across protein families.

Conclusions: Our results show that structurally conserved pockets are a common feature of protein families. The structural conservation of protein pockets, combined with other characteristics, can be exploited in drug discovery procedures, in particular for the selection of the most appropriate target protein and pocket for the design of drugs against entire protein families or subfamilies (*e.g.* for the development of broad-spectrum antimicrobials) or against a specific protein (*e.g.* in attempting to reduce side effects).

Background

Molecular processes in the living cell are coordinated and executed under tight regulation. Proteins play a fundamental role in almost all biological processes, and their overall activity is regulated at different levels [1]. At a first level, the concentration of a particular protein in the cell is regulated through its synthesis rate (gene expression) and its degradation rate. At another level, mechanisms act on the protein molecule itself through covalent modifications or non-covalent binding of small ligands or other molecules. These regulatory mechanisms are not only essential for the proper functioning of the molecular processes that maintain life, but are also responsible for cross-signaling and regulation processes between an organism and its environment.

Many metabolic enzymes, signalling proteins and transcription factors, among others, are regulated allosterically. Allosteric regulation has been studied for more than 50 years and it is considered the most powerful and common way to regulate protein activity [2]. However, for

* Correspondence: xavier.daura@uab.cat
[1] Institute of Biotechnology and Biomedicine (IBB), Universitat Autònoma de Barcelona (UAB), Bellaterra, E-08193, Spain
Full list of author information is available at the end of the article

most known cases of allosterism, the atomic details that explain the functional relationship between distant sites on the same protein molecule have not been elucidated [3,4].

Many pharmaceutical compounds act through allosteric regulation, as seen in the case of paclitaxel (Paxol), a cancer therapeutic drug that regulates tubulin polymerization allosterically [5,6]. Even though active sites represent the classic drug-target pocket (*e.g.* Aspirin and cyclooxygenase), allosteric sites present advantages over active sites in the context of drug design. Enzymatic activity usually involves charged transition states and the substrates are not always drug-like. Thus, orally active inhibitors that complement these sites can be very difficult to obtain. Moreover, allosteric sites may allow the discovery of not only novel drug-like inhibitors, but activators as well [2,3].

In this context, predicting allosteric sites computationally is of great interest. Allosteric sites have been predicted using structural information [7] and phylogeny [8]. Recently, methods have been developed in order to model or predict the relationship between allosteric and active sites [9-11]. These methods represent an important step forward in the understanding of allosterism. However, these studies are limited by the low quantity of readily available data on allosteric sites. As stated by Thornton and collaborators in their recent review [4], this is due in part to the lack of a formal database that organizes and stores knowledge on allosteric proteins and the corresponding mechanisms.

To unveil common patterns underlying allosterism, given that these exist, a large-scale study using structural and sequence data would be necessary. However, given the present scenario of scarce allosteric-site data, we decided to perform a large-scale analysis of protein ligand-binding pockets, as these represent potential locations of functional and allosteric or regulatory sites. Our approach is supported by the concept that besides naturally ocurring allosteric sites, serendipitous sites -having no natural ligand but effectively being an allosteric site given an appropriate ligand- may be of great pharmacological interest [2]. Examples of previously unknown allosteric sites discovered on already solved protein structures [12,13] support the idea that orphan or serendipitous allosteric sites exist which may lack a known natural effector, but provide an excellent opportunity for drug discovery approches such as virtual screening. Hardy and Wells also suggest that the large amount of 'crystallization artifacts' present in the Protein Data Bank (PDB) [14], such as ligands co-crystallized in unexpected binding sites, could hint the presence of previously unknown allosteric sites [2].

A large database of protein structures and associated small-molecule ligands is available [15] and has been used to predict ligand-binding sites by homology [16]. However, small-molecule ligands are not always easy to co-crystallize and we did not want to limit our study to only such cases. In this context, ligand-binding sites can be computationally predicted from structure alone with reasonable accuracy [17-20]. To our knowledge, ligand-binding pockets as predicted directly from structure [19] have not been studied or characterized at large-scale yet, even though they represent the potential location of yet unknown effectors [2].

Functional pockets in proteins have been previously characterized in terms of their flexibility [21,22], evolutionary conservation [21,23] and electrostatic potential [24] and these characteristics have been used to predict their presence and location in the protein structure [23]. Evolutionary conservation is a common characteristic of biologically functional sites. However, until now it has been exploited solely at the sequence level [23]. Although sequence and structural conservation correlate, structure is closer to function and may be conserved even in the lack of a sequence-level signal [25]. Despite this, to our knowledge, an approach based on the structural conservation of protein pockets has not been previously used. Here, we introduce a simple methodology to study pockets at the protein family level, consisting in the identification of pockets present in equivalent positions across different structures of the same protein family. To parameterize the method, we used protein pockets that matched known active sites, as these are well annotated [26,27]. Once parameterized, we applied the method to all protein structures available in the PDB [14], leading to the identification of protein pockets for thousands of different protein families [26]. Next, we compared the levels of structural conservation with other pocket characteristics estimated on the same protein families, such as sequence level conservation, backbone flexibility and electrostatic potential.

In the following sections we also discuss the results of this analysis for a small set of biological examples which illustrate the relevance of structural conservation in studying protein functional and regulatory sites. Finally, we perform an estimation of the amount of potentially paired regulatory and functional sites that may exist in the entire data set.

Results and Discussion

Initial structural data set

To acquire a large-scale perspective on the conservation of protein pockets, we gathered all available protein structures from the Protein Data Bank (PDB) [14]. We applied a set of filtering criteria to ensure the quality and relevance of the structural data before grouping the structures by protein families, as defined by the Protein families database (Pfam) [26]. To partially cope with the

inherent bias present in the PDB, where proteins tend to be over- or under-represented [28], we selected a set of representative structures for each protein family (see Methods). The final data set covered 4,258 different Pfam protein families and was composed of a total of 22,312 distinct protein structures (maximum 95% sequence identity), on which we predicted the location of 167,648 putative ligand-binding pockets by means of the LIGSITEcs program [19].

Identifying equivalent pockets across different protein structures

The first step to estimate the structural conservation of protein pockets was to identify those that appeared at equivalent positions in different structures of the same family. Briefly, for each protein family the pockets predicted for a representative set of structurally aligned proteins [29] were clustered following the approach described in the Methods section. The clustering method requires a threshold distance to select equivalent pockets across superimposed structures. After visual inspection of preliminary results, we observed that this parameter would be related to the structural fluctuation present in each protein family. We decided to use known active sites as a reference to define this parameter, as we were able to map a total of 8,046 pockets (covering 319 distinct protein families) to Pfam-annotated or predicted active-site residues unambiguously (see Methods). If the active site is well conserved across the whole protein family, an ideal clustering method would include all active sites of the different structures in the same cluster (true positives), without including any non-active site pockets (false positives). After benchmarking a range of different values (see Table 1), we defined the family-specific distance threshold to be 2.0 Å plus the average RMSD observed when superimposing the representative structures of the protein family. This approach showed a good compromise between true positives (including an average of 76.5% of all active sites) and false positives (including 8.95% of non-active site pockets), as shown in Table 1. For the families included in this study, the average value observed for the family-specific threshold was 4.5 Å.

Assessing the structural conservation of pockets on protein families

After parameterizing the clustering method using active sites as reference, we applied it to all protein families having at least 5 representative structures in the data set (a total of 1,128 protein families satisfied this requisite). We then analyzed the resulting clusters of pockets in terms of the percentage of representative structures covered by each cluster. A very well conserved pocket would be expected to appear in all representative structures of the protein family, *i.e.* 100% coverage. Thus, this coverage can be taken (and will be taken throughout this study) as a measure of the pocket's structural conservation within the protein family. This analysis was performed, for each family, for the first three clusters and for the cluster containing the largest amount of active sites (active-site cluster). The results are illustrated in Figure 1. Note that cluster ranking is based on average pocket size and coverage as described in the Methods, and that the active site cluster overlaps with the 1st, 2nd and 3rd clusters in 117, 38 and 13 families, respectively.

According to the histograms in this figure, the higher the coverage of the active site cluster the higher its frequency in the ensemble of families. This, which should be expected for pockets that are functionally relevant at the family level, is also applicable to the distribution of cluster 1 but not to those of clusters 2 and 3. Yet, the coverage distributions of clusters 2 and 3 suggest that they could be important at a sub-family level, remaining compatible

Table 1: Parameterization. Performance of the clustering algorithm when grouping known active sites.

Fixed distance value (Å)	% total active sites	% non active sites
0.0	59.60	4.46
1.0	69.89	5.31
2.0	76.47	8.95
3.0	79.13	13.67
4.0	82.78	17.30
5.0	83.50	23.04
6.0	86.71	27.90
7.0	87.43	34.68
8.0	87.62	41.97
9.0	88.24	49.59
10.0	89.29	56.87

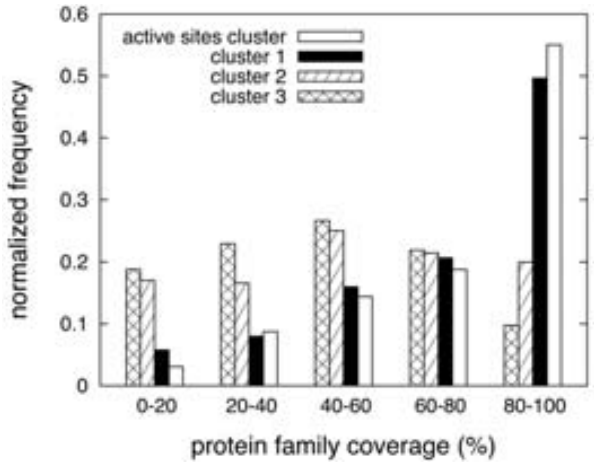

Figure 1 Structural conservation. Histograms displaying protein family coverage of different pocket clusters, for a total of 1,128 protein families with five or more representatives in the data set.

with an allosteric function which may have different faces within the same protein family. In global terms, the average coverage of the first cluster or most conserved pocket of the 1,128 protein families analyzed is 85%. Of these 1,128 protein families, 398 (35%) show at least one pocket cluster that covers 100% of the protein-family representatives, while 884 (78%) present a pocket cluster that covers at least 75% of the protein family.

These results show that for the majority of the protein families analyzed there is at least one pocket with high levels of structural conservation. We expected a high frequency of conserved pockets among enzymes, but not all protein families in the data set have been annotated with a biological activity that is related to a pocket in the protein structure. A structurally conserved pocket whose biological function has not been described is an optimal candidate for further computational and experimental analysis. For example, in the context of drug design and discovery, the information on whether a pocket on the target protein is structurally conserved or not may be useful when designing a wide-spectrum or a specific drug, respectively, and in choosing the appropriate ligand-binding site for virtual screening. Clearly a pocket that is very well conserved at the structural level may not necessarily have functional properties but be the consequence of structural restraints common across the protein family. Nevertheless, it may still be of interest to explore its possible exploitation as a serendipitous allosteric site for a therapeutic application [2].

Before further exploring these possibilities, we analyzed the degree of correlation between structural conservation and other properties often used for the characterization of protein pockets, such as evolutionary conservation at the sequence level [18,21,23], protein flexibility [21,22] and electrostatic potential [24]. These parameters may be useful in distinguishing pockets that are conserved because of their biological function from pockets that are conserved because of structural restraints.

Comparison with other pocket characteristics

Sequence conservation

Biologically relevant residues tend to be conserved at the sequence level [30]. In this context, the degree of conservation of the residues defining a protein cavity may be taken as a measure of the cavity's conservation [18-20]. The statistical significance of this measure can be then tested by comparing the levels of sequence conservation in the pocket and in the rest of the protein (see Methods).

Although structure is in general more conserved than sequence [25], the two characteristics are related. To analyze the relationship between sequence and structural conservation for pocket clusters across protein families, we quantified sequence conservation as the percentage of pockets in the cluster that are significantly conserved at the sequence level. The structural and sequence conservation values for clusters 1 to 3 of all protein families with at least five representative structures are compared in the left panel of Figure 2. It is shown that there is a relatively small correlation between sequence conservation and structural conservation of the pockets, with the highest density (62% of the population) at the 0-5% sequence-conservation end of the distribution. Yet sequence conservation is clearly peaked also at the 95-100% end (12% of the population), indicating that in general structurally conserved pockets (pocket clusters) may be well conserved at the sequence level, or not at all, leaving few cases in between.

One may argue that pocket clusters displaying high sequence and structural conservation may match biologically functional pockets across the protein family, while clusters displaying only structural conservation may often play a purely structural role. In relation to this, and giving the conservation percentages the right context, it should be noted that a large proportion of the protein families included in this analysis may not present a biological activity that is related to a particular pocket in the structure. Nevertheless, some pockets may be biologically relevant, despite a lack of sequence conservation. An example of this is given by L-lactate dehy-drogenase (LDH), for which the allosteric site [31] is very well conserved at the structural level (89.8%), but shows no signal of sequence conservation (0.0%) in this analysis. We describe the LDH case in further detail below.

As discussed above, pockets that are conserved at the structural level but have not been previously described as ligand-binding sites may be evaluated as potential orphan or serendipitous allosteric sites and targeted for drug discovery and design. The low sequence conservation we

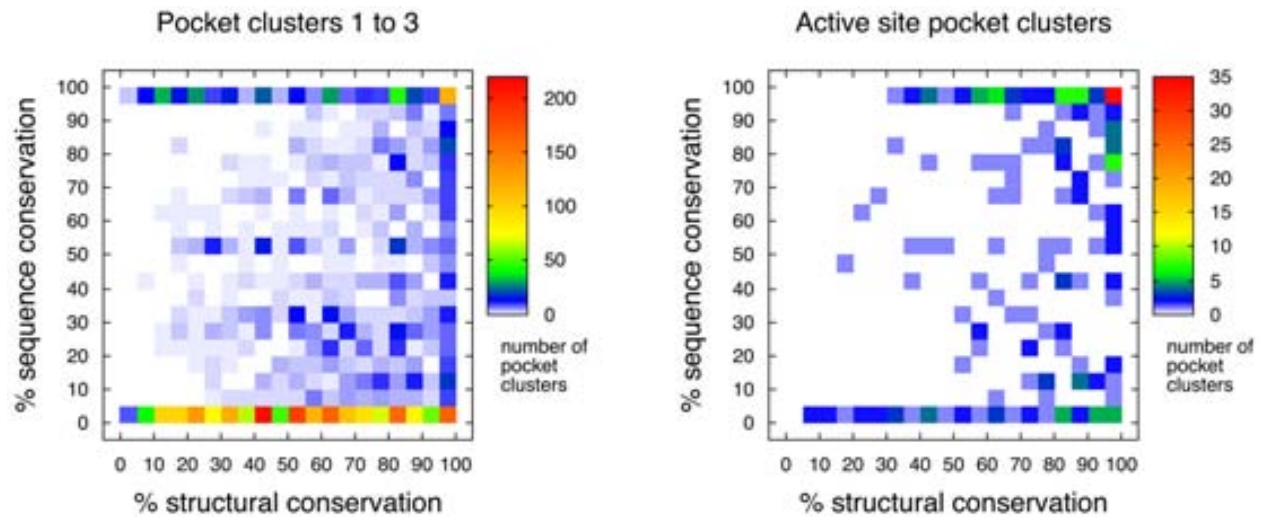

Figure 2 Sequence conservation. Two-dimensional histograms comparing structural and sequence conservation for pocket clusters of the different families in our data set. The left panel displays the values for pocket clusters 1 to 3 of 1,128 distinct protein families and the right panel shows the distribution for clusters containing the majority of active sites for 229 protein families. Only families with at least five representative structures are included in the analysis.

observed in many of the structurally conserved pockets indicates that even though the pocket is detected in the same location, the residues defining the pocket are not under direct evolutionary pressure and may vary in type. This variation in residue composition could aid the design of highly specific drugs that would bind only certain members of the protein family.

We performed the same distribution analysis for pocket clusters that included the largest amount of active sites in the corresponding protein families, with results displayed in the right panel of Figure 2. In this case, the largest population corresponds to high levels of both sequence and structural conservation, as expected, although numerous exceptions appear in this set as well. Some exceptions rise because not all members of a given Pfam protein family may be enzymatically active (*e.g.* Globin).

Note that the level of sequence conservation of each cluster is estimated from its member pockets and is independent of the weight of the cluster in the set of family-representative structures. A sequence conservation of 100% means that every single pocket in the cluster is significantly conserved at the sequence level, although the cluster may only cover half of the protein family, *i.e.* 50% structural conservation. This is also valid for the flexibility and electrostatic-potential analyses described below.

Flexibility

Protein function is fundamentally linked to dynamics. In this context, the properties of a protein's ligand-binding site are to an important extent a function of the site's flexibility, entropy being an essential component of the free energy of binding. Thus, relatively small changes in flexibility often have a large effect on ligand-binding affinites [32]. Moreover, some allosteric sites regulate protein function by modifying the protein's flexibility upon ligand binding [2,9,33].

Flexibility may be estimated on a residue basis from structural B factors. This has been previously used, for example, to show that active sites tend to be more rigid than the rest of the enzyme structure [21,22].

We analyzed the flexibility of residues forming part of pockets and determined if they showed significantly higher or lower values of flexibility than the rest of the protein's backbone, classifying them as 'flexible' or 'rigid', respectively, as described in the Methods section. We then compared the percentage of significantly rigid or flexible pockets found in the different pocket clusters with their structural conservation. The results are illustrated in Figure 3. These results show very few cases where pockets are significantly more flexible than the rest of the protein, and these few cases (clusters 1-3, left panel) tend to be poorly conserved in terms of structure. However, the structural conservation of a very flexible pocket would be probably hard to quantify by our method, given that a large degree of structural variation would to some extent impede its detection across different proteins of the same family. Even in average cases, member pockets of the same cluster can display large differences in shape and volume, as seen in the case of LDH described below.

The lower panels of Figure 3 show a wide distribution of significantly rigid pockets. This means that within a family, the levels of flexibility for the same pocket may differ from structure to structure considerably. These results were expected to a certain level, since flexibility

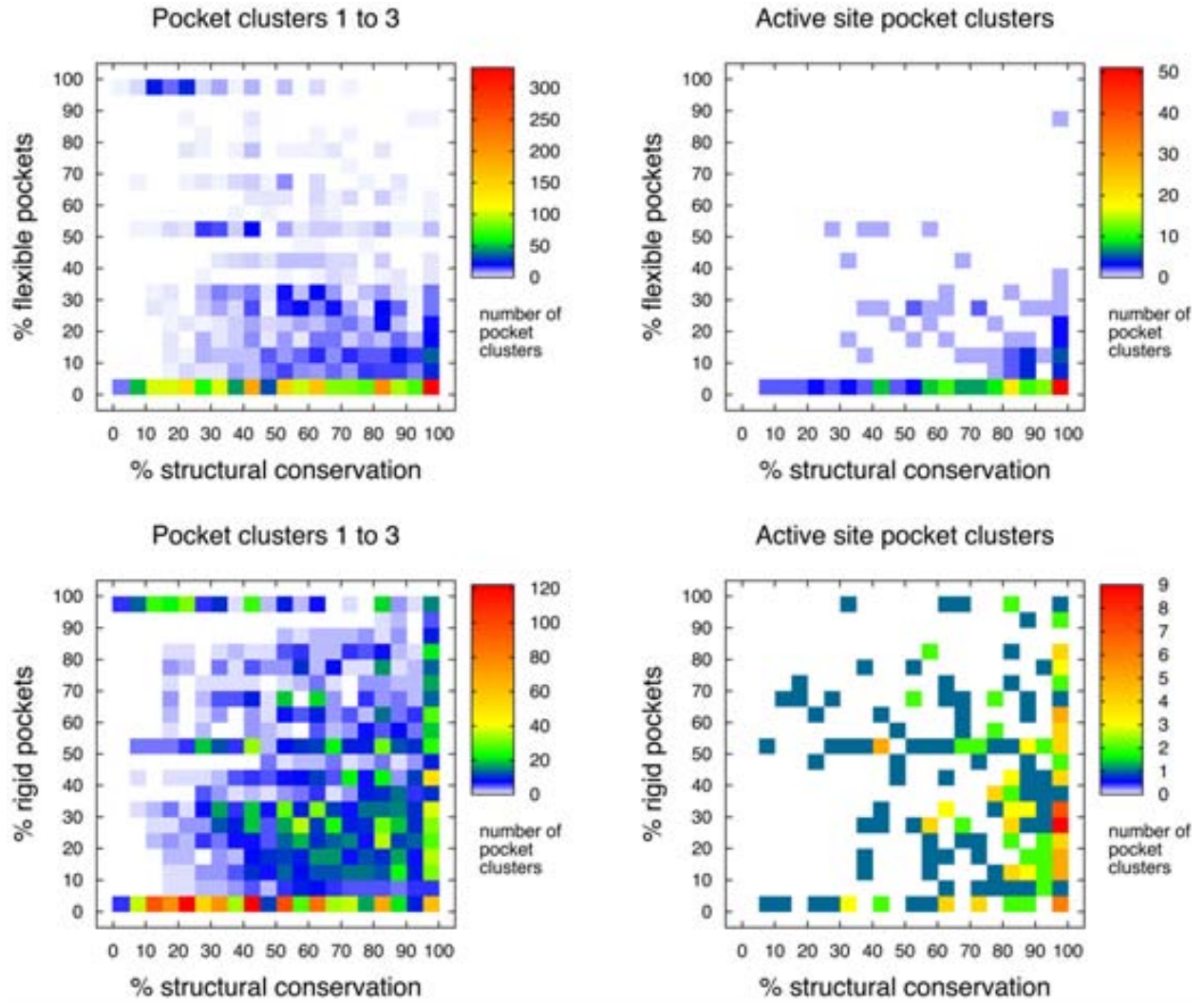

Figure 3 Flexibility. Two-dimensional histograms comparing structural conservation and flexibility for pocket clusters of the different families in the data set. The panels on the left display the values for pocket clusters 1 to 3 of 1,128 distinct protein families and the right panels show the distribution for clusters containing the majority of active sites for 229 protein families. The top panels show the percentage of pockets that were significantly flexible when compared to the rest of the structure, while the bottom panels show the percentage of pockets found to be significantly rigid (less flexible) than the rest of the corresponding protein structure. These plots only include protein families with at least five representative structures.

may vary under different experimental conditions of structure determination and it may as well be modified by the presence of bound ligands or other proteins [32]. In the case of active site pocket clusters, the lower-right panel of Figure 3 displays a stronger signal for structural conservation than for rigidity.

Protein flexibility is a major issue for ligand virtual screening and design [34]. Although key residues in active sites, such as those involved in catalysis, tend to be rigid [21,22], they coexist with regions of high flexibility, which are necessary to allow for ligand exchange. When searching for other possible ligand-binding sites for the screening and design of effector molecules, one will usually target pockets that are sufficently flexible that binding will not be blocked by high free-energy barriers (involving conformational rearrangements) but at the same time sufficiently rigid that computational docking will be reliable and that there will not be a sizable entropic penalty due to a potentialy large loss of flexibility upon ligand binding (which would need to be compensated enthalpically for effective binding). The analysis shown here might provide a basis to select structurally conserved pockets with specific flexibility properties.

Electrostatic potential

The electrostatic potential, as estimated by solving the Poisson-Boltzmann equation for protein structures with force-field-based charge distributions [35] has been previously used to characterize and predict enzymatic active

sites [24]. For each pocket in the data set we estimated the electrostatic potential at the pocket's center of mass as described in the Methods section and computed the average value over the pockets for each of the first three pocket clusters in every protein family. The combined distribution of average electrostatic potential and structural conservation of pocket clusters is shown in Figure 4. Clearly, this property does not correlate either with structural conservation. Most values cluster between -5 and 2.5 *kT/e*, even for active site pockets. However, this measure is probably the least conserved across the different pockets of a given cluster. In fact, for 42.4% of the pocket clusters included in the left panel of Figure 4, the standard deviation is larger than the average absolute value. It appears that the pocket's electrostatic potential, as estimated here, is largely protein specific, and that this measure is hard to extrapolate across the different proteins in a family. Nevertheless, the values of electrostatic potential can still be used in refining the selection of pockets for drug-screening procedures, given that drug-like ligands may be easier to find for more neutral sites than for strongly charged or polarized pockets [2]. These values could also be used to distinguish putative active sites from allosteric sites in the lack of proper annotation (see PIG-L below).

Biological examples

To complement the large-scale perspective presented above we analyzed a few protein families in more detail. The examples described below emphasize the relevance of structural conservation in the study of allosteric and functional protein pockets.

L-lactate dehydrogenase

L-lactate dehydrogenase (LDH) catalyzes the reduction of pyruvate by NADH to L-lactate in the last step of glycolysis. Certain bacterial LDHs, in contrast to their mammalian counterparts, display allosteric regulation by fructose 1,6-bisphosphate (FBP) [36]. Iwata and co-workers solved the structure of LDH ([PDB:1LTH]) in both active (R) and inactive (T) states, co-crystallized with the allosteric activator [31]. The Ldh_1_C domain in the R state (relaxed or active) of LDH is displayed in Figure 5A with the bound allosteric activator and pocket clusters 1, 2 and 4, as calculated for this family. Figures 5B-F show examples of distinct member pockets matching cluster 1 in the LDH protein family.

The active site in this protein matches pocket clusters 2 and 4. Both clusters are very well conserved at the sequence level with, respectively, 92.1% and 77.8% of the included pockets being significantly conserved. These active-site pockets are also well conserved at the structural level: cluster 2 appears on 75.5% of the representative structures while cluster 4 appears on 65.3% and, when considered together, at least one of them appears on 94% of the protein family. Interestingly, the pocket cluster with the highest structural conservation (cluster 1), corresponds to the allosteric site (Figure 5). In this case the allosteric cluster covers the majority of representative structures for this family (89.8%). However, the average sequence conservation signal is very low (-0.12) and we found none of the 51 pockets included in this cluster to be significantly conserved at the sequence level. This means that an evolutionary analysis based purely on sequence information would not find this site to be signif-

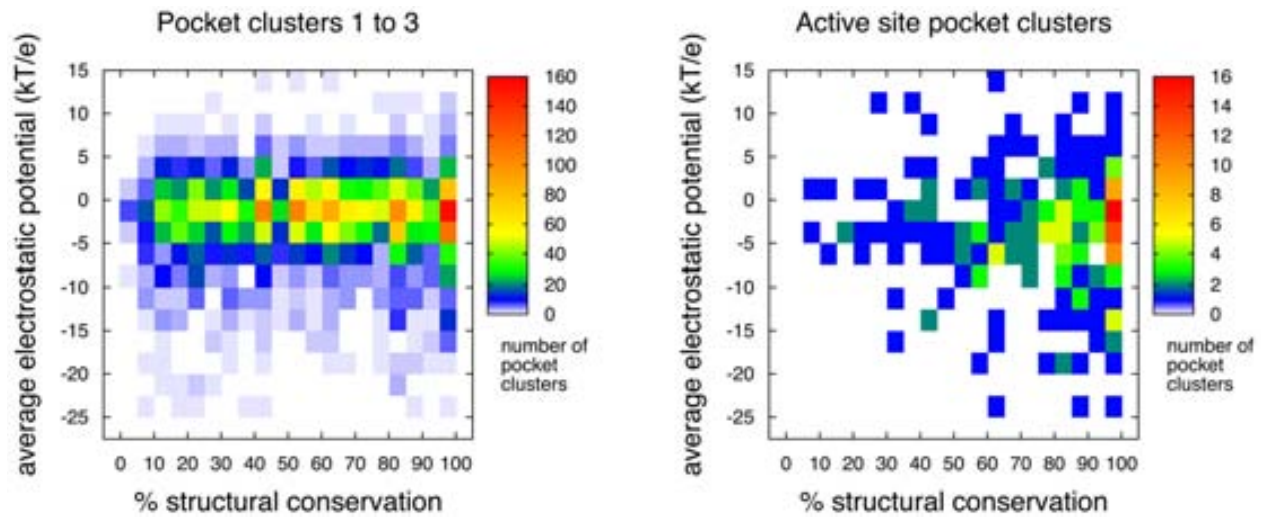

Figure 4 Electrostatic potential. Two-dimensional histograms comparing structural conservation and electrostatic potential for pocket clusters of the different families in the data set. The panels on the left display the values for pocket clusters 1 to 3 of 1,128 distinct protein families and the right panels show the distribution for clusters containing the majority of active sites for 229 protein families. These plots only include protein families with at least five representative structures.

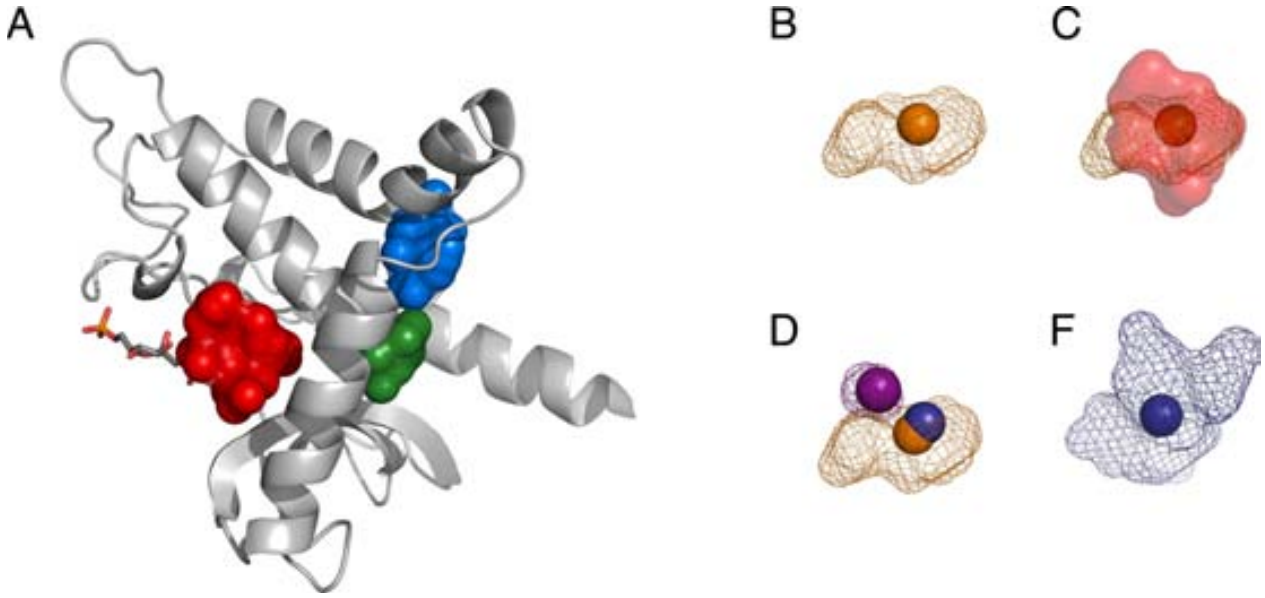

Figure 5 LDH pocket clusters. A) LDH structure ([PDB:1LTH], chain R, residues 150-317 which correspond to Pfam Ldh_1_C domain). Pocket clusters are displayed as the centers of mass of member pockets in 'spheres' representation (cluster 1: red, 2: green, 4: blue). The allosteric effector fructose 1,6-bisphosphate is displayed in 'sticks' representation next to cluster 1. Clusters 2 and 4 match the active site of LDH, while cluster 1 matches the allosteric site. B) Member pocket of cluster 1, found in structure [PDB:1SOV]. The center of mass of the pocket is displayed in 'spheres' representation while the volume of the actual cavity is shown in 'mesh' representation, both in orange color. C) The member pocket displayed in 'B' is shown in its relative position to the allosteric site cluster 1. Cluster 1 is shown in semi-transparent red 'surface' representation, similar as in A). D) The cluster 1 member pocket of [PDB:1SOV] displayed in orange in its relative position to other member pockets of the same cluster 1. The member pocket in purple was predicted on structure [PDB:1MLD], while the blue 'sphere' represents the center of mass of the member pocket found in yet another structure ([PDB:2A92]). F) Detail on the shape of the member pocket found in structure [PDB:2A92]. Structural orientations were kept constant (after protein backbone superimposition) to illustrate the varied shapes and volumes that member pockets of a single pocket cluster may display.

ically conserved, while the structure-based approach points it out as the most conserved pocket in this protein family.

The allosteric site cluster is remarkable in terms of flexibility as well, with 81.2% of included pockets being significantly rigid (see Methods). In the case of PDB entry [PDB:1LTH], there is a clear difference in the global flexibility values we calculated for the R (0.43) and T (-0.56) structures of the protein, corresponding to the active and inactive states, respectively. However, the allosteric site pocket shows consistently low values, -0.52 and -0.93 for R and T, respectively, with both pockets being significantly rigid according to the statistical test (p-values of 0.0008 and 0.0012, respectively). The active site in [PDB:1LTH] shows no significant differences in terms of flexibility when compared to the rest of the structure, although as expected, differences between the T and R states are also observed. The rigidity of the active site pockets through the whole family is not clear from the data, as 69.4% and 25.7% of the pockets are significantly rigid for clusters 2 and 4, respectively.

In the [PDB:1LTH] entry, the estimated electrostatic potential for the allosteric and active site have different values, with 1.22 and -8.09 *kT/e*, respectively, for the R state and similar values for the T state (1.36 and -5.73 *kT/e* respectively). This case matches the concept that active sites may bind more polar or charged molecules, while the allosteric site may bind more drug-like ligands [2]. When averaging these values over the corresponding member pockets, the standard deviation is close in magnitude to the values obtained, being 2.04 *kT/e* for the allosteric site cluster and -3.73 *kT/e* for the active site cluster 2. As discussed above, the electrostatic potential estimations tend to vary largely from structure to structure and thus are hard to extrapolate across the different proteins in a family.

Briefly, in this protein family we found the active site to match expected characteristics of biologically relevant pockets, being very well conserved both in terms of sequence and structure. The allosteric site, despite being very well conserved in terms of structure, does not appear to be conserved at the sequence level.

ADP Ribosylation factor 1

ADP-ribosylation factors (ARFs) are essential and ubiquitous in eukaryotes, being involved in vesicular transport and functioning as activators of phospholipase D and cholera toxin [37]. ARF activity is regulated by the binding and hydrolysis of GTP. The atomic structure [PDB:1HUR] shows the allosteric regulator bound to the protein [37], matching the position of cavity clusters 1

Table 2: Arf pocket clusters.

Cluster	structural conservation (%)	sequence conservation (%)	% flexible	% rigid
1	62.5	100.0	0.0	63.6
2	87.5	0.0	14.3	14.3
3	81.3	100.0	0.0	76.9
4	43.8	14.3	42.9	42.9
5	93.8	100.0	0.0	60.0

Properties of the principal pocket clusters of the Arf protein family. The structural conservation corresponds to the percentage of representative structures where a pocket of this cluster is present, the sequence conservation represents the percentage of pockets that are significantly conserved at the sequence level. The last two columns correspond to the percentages of significantly flexible or rigid pockets detected in the corresponding cluster.

and 3 as displayed in Figure 6. Both clusters matching the allosteric site show high levels of structural and sequence conservation as summarized in Table 2. These clusters also tend to be rigid, with clusters 1 and 3 having 63.6% and 76.9% of their pockets significantly rigid, respectively.

Cluster 1 matches the pyrophosphate group of GDP and the Mg ion as displayed in Figure 6. This cluster covers 62.5% of the representative structures of this protein family and we found all of the included pockets to be significantly conserved at the sequence level. The numbers for cluster 3 are similar, as 100% of its pockets are significantly conserved at the sequence level and are detected on 81.3% of the representative structures of this protein family. This cluster matches the pyrimidine-imidazole part of GDP as displayed in Figure 6.

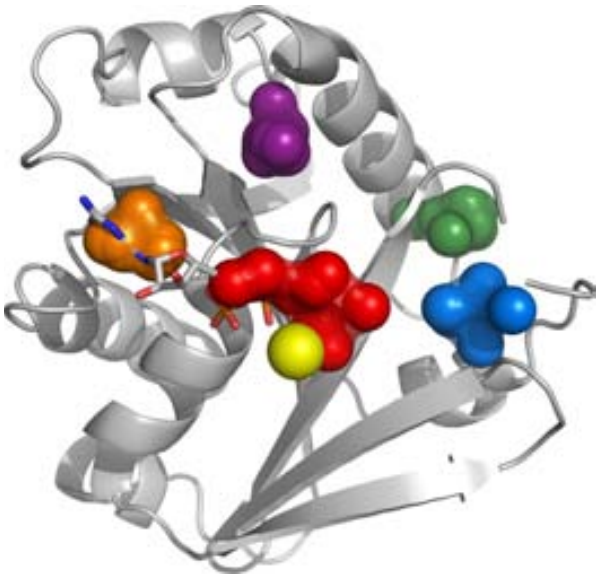

Figure 6 Arf pocket clusters. Human ADP-Ribosylation factor 1 structure ([PDB:1HUR], chain A which corresponds to Pfam Arf domain). The first five pocket clusters of this protein family are displayed by showing the centers of mass of the member pockets in 'spheres' representation (cluster 1: red, 2: green, 3: orange, 4: blue, 5: purple). The allosteric activator GDP is displayed in 'sticks' representation next to cluster 1, the Mg ion is colored yellow. Cluster 1 matches the pyrophosphate group of GDP and cluster 3 matches the pyrimidine-imidazole region of the allosteric ligand.

We do not know the biological function, if any, of the pockets represented by the rest of the clusters displayed in Figure 6. Although cluster 2 shows a high level of structural conservation covering 87.5% of the family, the pocket is not significantly conserved at the sequence level. Another interesting cluster is number 5, which appears on 93.8% of the structures and has a sequence conservation of 100%. This cluster is also rigid, with 60% of its pockets being significantly rigid, similarly to the clusters matching the allosteric site. It is interesting that cluster 4 shows almost half of its pockets to be significantly rigid and the other half to be significantly flexible as indicated by the values in Table 2. Initially we thought that this could be related to co-crystallized ligands affecting the flexibility of particular pockets through binding, but none of the analyzed structures presented a ligand in this position. We compared two structures corresponding to this family, namely PDB entry [PDB:1Z6X] (where the corresponding pocket is significantly flexible) and [PDB:1FZQ] (where the pocket was found to be sig-nificantly rigid). The rigid pocket in [PDB:1FZQ] was located next to an α-helix, while the same region in [PDB:1Z6X] lacked secondary structure presenting a loop-like conformation. It is remarkable that a pocket may be consistently found in two structures of the same family in a region with diverse secondary structure arrangements and levels of local backbone flexibility. Given that flexibility plays an important role in binding affinity [32], structurally conserved pockets may present distinct binding dynamics that can be exploited in the design of highly specific drugs.

Prediction of allosteric sites

The idea that yet undiscovered allosteric sites may be found in already solved structures has been mentioned in

a review by Hardy and Wells [2], in which they show various examples of previously undescribed allosteric sites found by serendipity. This concept, along with the few cases we examined in detail, prompted us to estimate the number of putative allosteric sites that may be found in the structural data set. We defined a simple estimator that consisted in scanning the data set for pairs of pocket clusters that are conserved at the structural level and are at least 8 Å appart (centroids distance). We performed this analysis on 1,128 protein families for which we had at least 5 representative structures. The results are presented in Table 3. A surprisingly large percentage of protein families (90.6%) presents at least one pair of pocket clusters that are both structurally conserved and at least 8 Å apart. A smaller fraction (54%) of protein families displays also a sequence conservation signal. If we also require one of the pockets to match active site annotations [38], the numbers are smaller but relatively large, *i.e.* for the total of 258 protein families in the database with an active site annotation, 207 (80.2%) present another structurally conserved pocket that is located at a distance of at least 8 Å.

For example, we found such a pair of pocket clusters on the structures corresponding to the PIG-L Pfam family. One of theses structures ([PDB:1Q7T]) corresponds to MshB from *Mycobacterium tuberculosis* and is considered a potential therapeutic target [39]. This protein lacks active site annotation in Pfam [38] or in the Catalytic Site Atlas [27]. However, the group of Baker *et al.* localized the active site when determining the protein's structure [39]. The active site predicted by Baker and co-workers matches cluster 2 in our predictions and is highly conserved both at the structure (100%) and sequence (80%) levels. Cluster 1 in this family is represented in all structures, although it shows no sign of sequence conservation. In the solved structure, it appears close to the location of a ligand referred to as a crystallization artifact [39]. Moreover, while the active-site-matching cluster shows a strong average electrostatic potential of -7.89 *kT/e*, cluster 1 presents a much more neutral average value of -1.23 *kT/e*. Cluster 3 presents also high levels of structural conservation (80%) and, in addition, of sequence level conservation (75%). Both clusters 1 and 3 would be interesting candidates for virtual screening in the search for an allosteric effector ligand.

On the large-scale perspective, the large amounts of putative allosteric sites we have counted may be an overestimation. Many of these cases may represent pockets that are merely the consequence of structural or functional requirements in other regions of the protein. It would be interesting to test for functional links between these regions [8,11]. However, many protein families do not necessarily perform functions that are associated to a certain pocket, such as the ADP Ribosylation factor discussed above. In these cases, it would not be necessary to find a pair of conserved pockets at a certain distance, since the regulatory site may be a pocket while the protein activity itself may not take place *via* such a structural feature.

Conclusions

We have developed a simple methodology to estimate the structural conservation of protein pockets, based on their position and size, and have applied it to the large amount of publicly available structural data, covering 4,258 distinct protein families and 22,312 protein structures. The analysis reported here indicates that the presence of structurally conserved pockets is a common feature across protein families and, in some cases, is accompanied by distinctive pocket characteristics in terms of sequence conservation, flexibility or electrostatic potential. Although correlations between the latter properties and structural conservation appear to be low in general, there is, as expected, a higher correlation between pocket

Table 3: Pairs of conserved distant pocket clusters, allosteric sites prediction.

Conservation thresholds (%)		# protein families			
structural	**sequence**	**total**	**sequence conserved***	**active site match**	**both**
50	50	1,022	614	207	165
50	75	1,022	484	207	136
75	50	434	264	93	76
75	75	434	207	93	61

Protein families for which we found a pair of structurally conserved pocket clusters at least 8 Å appart of each other. We used 50% and 75% as the structural conservation thresholds for each pocket cluster in each pair combined as shown in the first column with the sequence conservation threshold. *At least one of the pocket clusters was found to have at least 50% or 75% of its pockets significantly conserved at the sequence level. These results cover 1,128 protein families for which we had at least 5 representative structures.

structure and sequence conservation for active sites than for other types of annotated or putative ligand-binding pockets. Conserved pockets that lack annotation may represent new opportunities for drug discovery approaches such as virtual screening. In antimicrobial-discovery projects, for example, knowledge of the extent to which a putative ligand-binding site is present across a given protein family (*i.e.* orthologous proteins in a range of species or genuses) can be applied to the design of broad-spectrum drugs, as well as in dealing with drug toxicity, given that an ideal binding site for an antimicrobial would be present in proteins across many pathogenic species but not in a human homolog. In turn, additional pocket properties such as those considered here may be used for fine selection among pockets with the required level of structural conservation. Thus, we have shown specific examples illustrating that sequence conservation and electrostatic potential may be in some instances used to distinguish active sites from allosteric sites, the latter having a lower sequence-conservation signal and a more neutral potential in these examples. The data generated in this study is available upon request.

Methods

Structural data set

We organized the large number of structures available at the Protein Data Bank (PDB) [14] in protein families by querying all protein sequences, derived from the atomic coordinates of PDB entries, against the Pfam database (release 23.0) [26]. We performed all sequence-based queries by means of the HMMER software suite [40].

To ensure the quality of the structural data and its relevance to our study, we applied a set of filtering criteria. We evaluated the stereochemistry of protein structures using the PROCHECK program [41] and removed from our data set entries with a G-Factor value lower than -1.00. In the case of structures solved by crystallographic techniques, we also required a resolution of at least 3.0 Å. Entries not solved by Nuclear Magnetic Resonance (NMR) lacking a resolution value were discarded, independently of the technique used. We also discarded Pfam entries of type 'Motif' or 'Repeat', keeping only types 'Family' and 'Domain' that were assigned to structural regions spanning at least 30 residues.

All structures in our data set were parsed and organized according to the Pfam entry they were assigned to. However, given the bias present in the PDB [28], a protein family would be poorly represented by a redundant set containing all related structures. To partially remediate this, we clustered the structures in each Pfam entry according to sequence identity (95%) using complete-linkage hierarchical clustering. For each of the obtained clusters, the structure with the best resolution was chosen as the group representative.

Predicting ligand-binding pockets at the protein-family level

To compare the spatial positions of potential ligand-binding pockets in different structures of the same protein family, we first superimposed the representative structures to a common reference by means of the MAMMOTH program [29]. The protein with the longest sequence in the family was taken as reference for the structural fit. If length alone failed to select a single reference structure, we used resolution as the second selection parameter.

We proceeded to predict putative ligand-binding pockets on the fitted structures using the LIGSITEcs program [19]. Residues were assigned to pockets according to a common distance criterion, which includes all residues within 8 Å of the pre-calculated pocket's center of mass [19]. Note that the standalone version of the LIGSITEcs program we used is different from the LIGSITEcsc version also mentioned in [19], as the former does not incorporate residue conservation as a parameter.

At this point, for each protein family we had a group of superimposed representative protein structures for which the location of putative ligand-binding pockets had been predicted. We then grouped together pockets found consistently in the same position in different representative structures of the same protein family using the clustering method described below. The calculated clusters were finally ranked according to the average size of their member pockets and the percentage number of family representatives featuring the pocket (coverage).

Clustering of pockets

We clustered putative ligand-binding pockets found in different representative structures, previously superimposed, of a protein family using a modified version of a previously described clustering algorithm [42]. In this case, the elements to cluster are the centers of mass of pockets and the metric used to define distances between elements is the Euclidean distance. The clustering algorithm makes no distinction between pockets belonging to different, superimposed structures or to the same structure. Given that the degree of structural diversity among representatives of a protein family varies across protein families, the threshold to the metric for the definition of neighbor elements was chosen to be family specific, as described in the Results section. Unlike the previous implementation of the algorithm, cluster selection is not made by straight neighbor counting but by the sum of neighbor pocket sizes, as predicted by LIGSITEcs [19].

The algorithm outline is as follows: (1) the center of mass of each predicted pocket (element) in the set of representative structures of a protein family is assigned a parameter corresponding to the size of the pocket; (2) the Euclidean distance between every pair of elements is cal-

culated; (3) a threshold distance is applied to identify the neighbors of each element in the family; (4) each element is scored by the sum of its size parameter with those of all its neighbors; (5) the element with the highest score is chosen as the center of a cluster, which is formed by all its neighbors; (6) the members of the selected cluster are removed from the pool of elements and the procedure is repeated until the pool is empty; (7) clusters are ranked according to their score, calculated as

$$Score_c = Str_c * \frac{\sum_{i=1}^{n_c} size_i}{n_c} \quad (1)$$

where $Score_c$ is the cluster's score, n_c is the number of pockets in the cluster, $size_i$ is the size of member pocket i and Str_c is the cluster's coverage of the protein family or structural conservation, as described in Results. Str_c is computed by

$$Str_c = 100 * \frac{n_r}{m} \quad (2)$$

where m is the total number of representative structures for the corresponding protein family and n_r is the number of representative structures with at least one pocket present in the cluster.

Sequence conservation

There are multiple methods to estimate sequence conservation starting from a multiple sequence alignment (MSA) [30]. We estimated the degree of positional conservation for every residue in our structure data set by the following procedure: (1) We aligned all sequences in the Pfam 'full' MSA [26] using the HMMALIGN program [40] and the corresponding HMM profile. (2) We computed the entropy of each position of the alignment by means of the AL2CO program [30], activating the program option that weights each sequence to partially compensate MSA composition bias [43]. (3) The entropy values for each position in the MSA were inverted (higher score means higher degree of conservation) and normalized by the observed standard deviation. (4) For each Pfam entry we stored the conservation scores obtained using the HMM profile positions as a reference. (5) We aligned each of our structures to the corresponding HMM profile and assigned the previously computed conservation scores to each residue.

To test if a pocket was significantly conserved at the sequence level, we compared the sequence conservation values obtained for all the residues within 8 Å of the center of mass of the pocket with those for all residues in the structure by applying the Wilcoxon-Mann-Whitney non-parametric test. We defined as significant those cases where the p-value <= 0.05.

Electrostatic potential

We estimated the electrostatic protential at the center of mass of the protein ligand-binding pockets by means of the DELPHI software suite [35], which provides finite-difference solutions to the Poisson-Boltzmann equation. First, we added hydrogen atoms to each structure in our data set using the REDUCE program [44], then proceeded to estimate the electrostatic potential by means of the DELPHI program, with default parameters.

Protein backbone flexibility

We estimated protein backbone flexibility from normalized B factors as previously described [21,22,45]. For each Cα atom in the structure, the flexibility is equivalent to its B factor after normalization by equation 3.

$$B' = \frac{(B - <B>)}{\sigma(B)} \quad (3)$$

where $< B >$ is the average over all Cα atoms in the structure and $\sigma(B)$ is the standard deviation. We then define a residue's relative backbone flexibility as the B' value of its Cα.

For NMR entries, which lack B factors, we calculated the root-mean-square fluctuation (RMSF) of each Cα atom over the ensemble of NMR models [46]. RMSF values may be in turn converted to pseudo B factors [47], by

$$B = \frac{8*\pi^2}{3} * RMSF^2 \quad (4)$$

We tested for pockets that differed significantly from the complete structures in terms of their flexibility. For each pocket in each structure, we compared the values obtained for residues within 8 Å of the center of mass of the pocket to those for all residues in the structure by applying the Wilcoxon-Mann-Whitney non-parametric test. If the values for the pocket were significantly higher, we marked the pocket as 'flexible' and if they were significantly lower, we marked the pocket as 'rigid'. We defined as significant those cases where the p-value <= 0.05.

Mapping active site residues to pockets

Active-site residue predictions by Pfam [38] usually involve between one and three residues. We combined this sequence level information with the structural prediction of pockets on protein structures by mapping active sites to predicted pockets. For each structure, we marked as active site the pocket that included the majority of predicted active site residues.

In many cases, more than a single pocket contained one or more active site residues. We marked these cases as ambiguous, to distinguish them from cases where the mapping was unambiguous, *i.e.* all active site residues contained in a single pocket.

Authors' contributions

XD and AP conceived the study and wrote the manuscript. AP carried out the computational work. All authors read and approved the final manuscript.

Acknowledgements

We thank Juan Cedano, Francisco Melo, Amelie Stein, Roberto Mosca, Roland Pache and Dunja Urosev for helpful comments and technical advice. This project is supported by funding under the Seventh Research Framework Programme of the European Union (ref. HEALTH-F3-2009-223101). AP acknowledges the FPU Scholarship from MICINN, Spanish Gov.

Author Details

[1]Institute of Biotechnology and Biomedicine (IBB), Universitat Autònoma de Barcelona (UAB), Bellaterra, E-08193, Spain and [2]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, E-08010, Spain

Received: 2 January 2010 Accepted: 31 March 2010
Published: 31 March 2010

References

1. Pardee AB: **Regulatory molecular biology.** *Cell Cycle* 2006, **5(8):**846-852.
2. Hardy JA, Wells JA: **Searching for new allosteric sites in enzymes.** *Curr Opin Struct Biol* 2004, **14(6):**706-715.
3. Cui Q, Karplus M: **Allostery and cooperativity revisited.** *Protein Sci* 2008, **17(8):**1295-1307.
4. Laskowski RA, Gerick F, Thornton JM: **The structural basis of allosteric regulation in proteins.** *FEBS Lett* 2009, **583(11):**1692-1698.
5. Han Y, Malak H, Chaudhary AG, Chordia MD, Kingston DG, Bane S: **Distances between the paclitaxel, colchicine, and exchangeable GTP binding sites on tubulin.** *Biochemistry* 1998, **37(19):**6636-6644.
6. Mitra A, Sept D: **Taxol allosterically alters the dynamics of the tubulin dimer and increases the flexibility of microtubules.** *Biophys J* 2008, **95(7):**3252-3258.
7. Freire E: **Can allosteric regulation be predicted from structure?** *Proc Natl Acad Sci USA* 2000, **97(22):**11680-11682.
8. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286(5438):**295-299.
9. Balabin IA, Yang W, Beratan DN: **Coarse-grained modeling of allosteric regulation in protein receptors.** *Proc Natl Acad Sci USA* 2009, **106(34):**14253-14258.
10. Kidd BA, Baker D, Thomas WE: **Computation of conformational coupling in allosteric proteins.** *PLoS Comput Biol* 2009, **5(8):**e1000484.
11. Daily MD, Gray JJ: **Allosteric communication occurs via networks of tertiary and quaternary motions in proteins.** *PLoS Comput Biol* 2009, **5(2):**e1000293.
12. Hardy JA, Lam J, Nguyen JT, O'Brien T, Wells JA: **Discovery of an allosteric site in the caspases.** *Proc Natl Acad Sci USA* 2004, **101(34):**12461-12466.
13. Morita K, Kawana K, Sodeyama M, Shimomura I, Kagechika H, Makishima M: **Selective allosteric ligand activation of the retinoid X receptor heterodimers of NGFI-B and Nurr1.** *Biochem Pharmacol* 2005, **71(1-2):**98-107.
14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
15. Stuart AC, Ilyin VA, Sali A: **LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures.** *Bioinformatics* 2002, **18:**200-201.
16. Marti-Renom MA, Rossi A, Al-Shahrour F, Davis FP, Pieper U, Dopazo J, Sali A: **The AnnoLite and AnnoLyze programs for comparative annotation of protein structures.** *BMC Bioinformatics* 2007, **8(Suppl 4):**S4.
17. Brady GP, Stouten PF: **Fast prediction and visualization of protein binding pockets with PASS.** *J Comput Aided Mol Des* 2000, **14(4):**383-401.
18. 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.
19. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.** *BMC Struct Biol* 2006, **6:**19.
20. Huang B: **MetaPocket: a meta approach to improve protein ligand binding site prediction.** *OMICS* 2009, **13(4):**325-330.
21. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: **Analysis of catalytic residues in enzyme active sites.** *J Mol Biol* 2002, **324:**105-121.
22. Yuan Z, Zhao J, Wang ZX: **Flexibility analysis of enzyme active sites by crystallographic temperature factors.** *Protein Eng* 2003, **16(2):**109-114.
23. Nicola G, Smith CA, Abagyan R: **New method for the assessment of all drug-like pockets across a structural genome.** *J Comput Biol* 2008, **15(3):**231-240.
24. Bate P, Warwicker J: **Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods.** *J Mol Biol* 2004, **340(2):**263-276.
25. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5(4):**823-826.
26. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008:D281-D288.
27. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004:D129-D133.
28. Xie L, Bourne PE: **Functional coverage of the human genome by existing structures, structural genomics targets, and homology models.** *PLoS Comput Biol* 2005, **1(3):**e31.
29. Ortiz AR, Strauss CEM, Olmea O: **MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.** *Protein Sci* 2002, **11(11):**2606-2621.
30. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment.** *Bioinformatics* 2001, **17(8):**700-712.
31. Iwata S, Kamata K, Yoshida S, Minowa T, Ohta T: **T and R states in the crystals of bacterial L-lactate dehydrogenase reveal the mechanism for allosteric control.** *Nat Struct Biol* 1994, **1(3):**176-185.
32. Teague SJ: **Implications of protein flexibility for drug discovery.** *Nat Rev Drug Discov* 2003, **2(7):**527-541.
33. Daily MD, Gray JJ: **Local motions in a benchmark of allosteric proteins.** *Proteins* 2007, **67(2):**385-399.
34. Carlson HA: **Protein flexibility and drug design: how to hit a moving target.** *Curr Opin Chem Biol* 2002, **6(4):**447-452.
35. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B: **Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects.** *J Comput Chem* 2002, **23:**128-137.
36. Garvie EI: **Bacterial lactate dehydrogenases.** *Microbiol Rev* 1980, **44:**106-139.
37. Amor JC, Harrison DH, Kahn RA, Ringe D: **Structure of the human ADP-ribosylation factor 1 complexed with GDP.** *Nature* 1994, **372(6507):**704-708.
38. Mistry J, Bateman A, Finn RD: **Predicting active site residue annotations in the Pfam database.** *BMC Bioinformatics* 2007, **8:**298.
39. McCarthy AA, Peterson NA, Knijff R, Baker EN: **Crystal structure of MshB from Mycobacterium tuberculosis, a deacetylase involved in mycothiol biosynthesis.** *J Mol Biol* 2004, **335(4):**1131-1141.
40. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14(9):**755-763.
41. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM: **AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.** *J Biomol NMR* 1996, **8(4):**477-486.
42. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE: **Peptide Folding: When Simulation Meets Experiment.** *Angewandte Chemie International Edition* 1999, **38(1-2):**236-240.
43. Henikoff S, Henikoff JG: **Position-based sequence weights.** *J Mol Biol* 1994, **243(4):**574-578.
44. Word JM, Lovell SC, Richardson JS, Richardson DC: **Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.** *J Mol Biol* 1999, **285(4):**1735-1747.

45. Carugo O, Argos P: **Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors.** *Proteins* 1998, **31(2):**201-213.
46. Wilmanns M, Nilges M: **Molecular replacement with NMR models using distance-derived pseudo B factors.** *Acta Crystallogr D Biol Crystallogr* 1996, **52(Pt 5):**973-982.
47. Willis BTM, Pryor AW: *Thermal vibrations in crystallography* Cambridge University Press; 1975.

doi: 10.1186/1472-6807-10-9
Cite this article as: Panjkovich and Daura, Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery *BMC Structural Biology* 2010, **10**:9

3.3 Exploiting protein flexibility to predict allosteric sites

Once we finished the large-scale study, we started to compile a data-set of structurally known allosteric proteins. This task was not trivial because the allosteric term is used quite freely across the literature (particularly in crystallographic articles) and we decided to skip protein structures for which there was no experimental proof or measure of a significant change in activity upon ligand binding. Moreover, authors may refer to protein-protein interactions, peptides and DNA as allosteric ligands and the definition is further extended to include phosphorylation and other post-translational modifications as allosteric effectors. During this work we preferred to focus on small-molecule ligands, given that these should be more straighforward to mimick using drug-like molecules. Following the same reasoning, we also skipped single atom effectors such as Calcium and other ions. After spending months parsing the literature we built a database containing structurally known allosteric proteins. However, our database was denied publication because a few weeks after submission, the AlloSteric Database (ASD) (Huang *et al.*, 2011) was published. We continued to work by parsing and incorporating the data available at ASD. In total we gathered 91 distinct allosteric protein structures for which the location and nature of the allosteric ligand was known.

As explained in the introduction, changes in protein flexibility caused by the binding of the allosteric effector is considered one of the most common mechanisms behind allostery. We attempted to capture such phenomena by utilizing normal mode analysis (NMA), which is a simplified approach to study protein dynamics. The work described in this second article (Panjkovich and Daura, 2012) consists in two major experiments. In the first one, we assessed the relevance of allosteric effector presence on the overall protein dynamics. For 70% of our data set, the presence of the allosteric ligand exerted a significant difference on the overall protein flexibility according to our analysis. We were surprised that a simplified computational approach could detect an allosteric effect on the majority of the cases. It is important to note that not all allosteric mechanisms are necessarily expected to function through a shift in overall protein dynamics. For example, the allosteric ligand may enhance the connecting interfaces of monomers to allow the formation of a higher order active complex as in the case of GTP cyclohydrolase stimulatory complex (Maita *et al.*, 2002).

The second part of the article describes how we combined our previous findings into a more applied goal. We tested the possibility of predicting the location and presence of allosteric sites by querying the putative ligand-binding sites on the protein structure through NMA. Basically, if a simplified representation of the ligand in that particular pocket exerted an overall effect on the flexibility of the protein (as compared to the unbounded or *apo* state) this would indicate a high chance of the pocket being an allosteric site, in a similar way as it was done previously

by others (Ming and Wall, 2006; Mitternacht and Berezovsky, 2011). We conducted this test on 58 proteins (we skipped those where the initial binding pocket prediction done by LIGSITE (Huang and Schroeder, 2006) did not match the allosteric site) obtaining 0.76 accuracy and doubling the positive predictive value (PPV) of a random approach. The performance increased when we combined this approach with our previously described structural conservation measure. When we used a stricter set of parameters (limiting the order of pockets analyzed) we obtained 0.65 PPV and 0.89 accuracy; however, the method is far from being a perfect predictor since sensitivity fell to 0.22. The detailed statistics of the predictions are displayed in Table 1 of the second article (Panjkovich and Daura, 2012).

Following the pattern established in the first article, we analyzed the application of the approach on a few proteins in more detail. HIV reverse transcriptase illustrates an intuitive case of correct allosteric-site prediction. Our method succeeds at locating the allosteric site in the hinge of the structure. Such a location is ideal to understand the underlying principle of the approach. The hinge connects two globular domains under fluctuation. A ligand able to bind this region will most probably perturb the dynamics of this protein by creating new interactions between the domains, thus affecting their rate of movement and in turn the activity rate. The details about the prediction on this structure can be seen in Figure 5 of the second included article (Panjkovich and Daura, 2012).

Out of the four cases analyzed in detail in the article, L-lactate dehydrogenase (LDH) is particulary interesting. The protein analyzed belongs to *Bifidobacterium longum* and illustrates previously mentioned advantages of allosteric drug-design over classic orthosteric approaches. This protein is known to be allosterically regulated in Gram-positive organisms at the predicted site. However, the homolog protein present in higher order organisms, including human, is not sensitive to allosteric regulation. Furthermore, this protein was used as an example in the previous article as well (Panjkovich and Daura, 2010) because the allosteric site shows a high degree of structural conservation but lacks any signal of conservation at the sequence level. As such, this allosteric site poses an ideal target for antibacterial development. A drug targeting this site would not bind the human homolog protein because the actual residues forming it differ significantly, which would drastically diminish any toxic or side effects on the patient. Moreover, even if the drug would bind the same site on the human protein, it would hardly exert any regulatory effect since the human homolog is not sensible to allosteric regulation. In line with this, our method predicted that binding the same pocket on the human LDH would not have a significant effect. Further details are available in the second article and the protein structure is displayed in its Figure 6 (Panjkovich and Daura, 2012). Overall, it is remarkable that a method based on a coarse-grained approximation to protein dynamics and a measure of structural conservation achieved such precise results. Furthermore, the method is quick in its

execution, a standard run on a monomeric protein of 300 residues will take around 2 minutes, while a large protein (500 residues) will be done in 6 minutes. Even though calculation times increase exponentially with the number of residues, considerably larger systems can be processed within reasonable time; for example, a protein of 1000 residues will be analyzed within 45 minutes. This makes the approach ideal for scanning large sets of protein structures, particularly in the case of structural genomics projects where protein structures are solved for which function and other biological information is unknown (Laskowski and Thornton, 2008).

Panjkovich and Daura *BMC Bioinformatics* 2012, **13**:273
http://www.biomedcentral.com/1471-2105/13/273

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Exploiting protein flexibility to predict the location of allosteric sites

Alejandro Panjkovich[1] and Xavier Daura[1,2*]

Abstract

Background: Allostery is one of the most powerful and common ways of regulation of protein activity. However, for most allosteric proteins identified to date the mechanistic details of allosteric modulation are not yet well understood. Uncovering common mechanistic patterns underlying allostery would allow not only a better academic understanding of the phenomena, but it would also streamline the design of novel therapeutic solutions. This relatively unexplored therapeutic potential and the putative advantages of allosteric drugs over classical active-site inhibitors fuel the attention allosteric-drug research is receiving at present. A first step to harness the regulatory potential and versatility of allosteric sites, in the context of drug-discovery and design, would be to detect or predict their presence and location. In this article, we describe a simple computational approach, based on the effect allosteric ligands exert on protein flexibility upon binding, to predict the existence and position of allosteric sites on a given protein structure.

Results: By querying the literature and a recently available database of allosteric sites, we gathered 213 allosteric proteins with structural information that we further filtered into a non-redundant set of 91 proteins. We performed normal-mode analysis and observed significant changes in protein flexibility upon allosteric-ligand binding in 70% of the cases. These results agree with the current view that allosteric mechanisms are in many cases governed by changes in protein dynamics caused by ligand binding. Furthermore, we implemented an approach that achieves 65% positive predictive value in identifying allosteric sites within the set of predicted cavities of a protein (stricter parameters set, 0.22 sensitivity), by combining the current analysis on dynamics with previous results on structural conservation of allosteric sites. We also analyzed four biological examples in detail, revealing that this simple coarse-grained methodology is able to capture the effects triggered by allosteric ligands already described in the literature.

Conclusions: We introduce a simple computational approach to predict the presence and position of allosteric sites in a protein based on the analysis of changes in protein normal modes upon the binding of a coarse-grained ligand at predicted cavities. Its performance has been demonstrated using a newly curated non-redundant set of 91 proteins with reported allosteric properties. The software developed in this work is available upon request from the authors.

Background

Proteins can be regarded as the functional building blocks of life, carrying out and coordinating almost all biological processes. Tight regulation of these processes is fundamental in all kingdoms of life and allostery represents one of the most commmon and powerful means of modulating protein activity [1]. Allostery can be defined as the regulation of a protein's function by binding of an effector molecule at a site which is not the active site. Its relevance was emphasized decades ago by Jacques Monod, when he referred to allosteric regulation as the 'second secret of life', second only to the genetic code [2]. Even though allostery and its often intrincate nature have captured the interest of researchers since the initial discoveries more than half a century ago (for a review see [3]), most allosteric mechanisms are still not completely understood [1]. At present, allosteric phenomena are being intensively studied for their potential as target mechanisms for the development of new classes of therapeutics [4].

Expanding drug-design through allostery opens up an unexplored territory of novel potential therapeutic solutions, beyond what has been already covered by the classic, active-site oriented drug-development approach.

*Correspondence: xavier.daura@uab.cat
[1]Institute of Biotechnology and Biomedicine (IBB), Universitat Autònoma de Barcelona (UAB), 08193 Cerdanyola del Vallès, Spain
[2]Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain

An important factor fueling interest in allosteric drugs consists in their characteristic advantages compared to traditional active-site inhibitors. For example, allosteric sites tend to be under lower sequence-conservation pressure than active sites, facilitating the design of highly specific drugs and reducing the risks of toxicity or side-effects [5-7]. To explain this briefly, if the pathogen's active site is very well conserved in nature it may share important structural features with the human homologue, which could be then bound and inhibited as well by the antimicrobial drug causing toxic side-effects on the patient. Thus, lower levels of evolutionary conservation at ligand-binding sites may allow for more selective drugs. Furthermore, allosteric drugs may not only inhibit but also increase target-protein activity, enabling novel therapeutic possibilites as seen for example in the activation of glucokinase by allosteric drugs, a potential treatment for type 2 diabetes mellitus [8,9]. On the same line, traditional drugs may be complemented by allosteric effectors, as observed in the case of aminoglycoside phosphotransferase where a previously unknown binding site could be exploited to allosterically counteract antibiotic resistance [10].

However, the field of allosteric-drug design is rather young and the amount of allosteric drugs known today is still marginal [7]. For example, at the time of this writing a query in DrugBank [11] for the term 'allosteric' returns 7 results, while 'inhibition' returns 483 entries. This may be in part due to the intrinsic difficulties in understanding allosteric mechanisms and to the lack of systematic studies on the topic [12]. Only recently the first initiative to store and organize information on allosteric cases has surfaced in the form of the AlloSteric Database (ASD) [13]. By browsing ASD it becomes apparent that part of the difficulty in studying allosteric systems lies in the large degree of variety found among them, as there are many ways in which protein activity can be affected allosterically [12,14]. A textbook example is the one provided by glucose-induced glucokinase, in which the ligand triggers a conformational change allowing the active site to become functional [15]. In other cases the presence of the allosteric ligand triggers the formation of the biologically active protein complex (*e.g.* GTP cyclohydrolase stimulatory complex [16]). A protein illustrating the variety and complexity allosteric mechanisms may reach is ribonucleotide reductase. This protein presents two different allosteric sites: one affects the enzyme catalytic rate and the other alters its specificity allowing the enzyme to switch substrates [17]. Furthermore, allosteric signals may also propagate solely by altering protein dynamics, without a detectable conformational change [18,19].

In the context of such diversity, unveiling common patterns beneath allosteric phenomena could increase their potential for therapeutic exploitation, stimulating the design of allosteric drugs. We postulate that the first step in such a procedure would be to computationally detect or predict the presence and location of protein allosteric sites, to allow further focusing of drug-screening processes on selected protein targets down the pipeline. The algorithm should be able to pinpoint which proteins are sensible to allosteric regulation. However, if as already suggested any dynamic protein has the potential to be regulated allosterically [20], then the method should indicate the location of putative allosteric sites on the protein. Based solely on sequence, it would be very hard to predict the location of allosteric sites as it has been done by homology on active sites [21,22], because the evolutionary pressure for sequence conservation on allosteric sites is generally much lower and harder to detect, if at all present [3,23].

Until now, much of the research in the field has been focusing on the conformational changes induced by allosteric signals. The group of Jeffrey Gray studied conformational changes upon allosteric activation [24] and expanded this research by analysing the networks of quaternary and tertiary motions on which allosteric communication relies [25]. Following a similar line, a very interesting and thorough study was published where different parameters were interrogated in terms of their potential to indicate which protein residues are involved in transmitting the allosteric signal, on the basis of experimental mutation data [26]. The results from these analyses aim at defining the particular pathway of residues that mediate the allosteric communication. However, other authors have argued that this may not be the case *in vivo*, where multiple effector sites may be present on the protein acting through multiple signaling pathways [27]. In general, recent studies aggree in the idea that allostery is mainly a thermodynamic process and among the different protein properties that are involved in allosteric phenomena, flexibility (*i.e.* protein dynamics) stands out as the most significant one [3,28-31].

Following this line of thought, Ming and Wall developed a theoretical framework to study allosteric effects by comparing the dynamics of bound and unbound protein-ligand pairs [32]. They further refined their methodology and tested its ability to predict functional ligand-binding sites (not necessarily linked to allosterism) on a set of 305 protein-ligand complexes of known structure [33]. Very recently, two other approaches partially aiming at predicting allosteric sites have been published by Mitternacht and coworkers. In a first article they describe a geometric measure that helps at locating biologically functional ligand-binding sites, while a second one describes a more elaborate measure called 'binding leverage', related to protein dynamics, which appears useful at locating biologically relevant binding sites including allosteric sites. They tested this last feature on 15 allosteric proteins

[34,35], observing different results for specific proteins and concluding that regulatory sites may be identified without previous experimental knowledge on conformational changes. However, these studies were not completely focused on allosteric sites and did not benefit from the larger data set now available at ASD [13].

Even though the previously cited articles represent an important step forward in the understanding of allostery, we consider that further research is needed if allosteric sites are to be predicted with the same coverage and precision as active sites [21,22]. The first thing we did in this direction was to integrate more than one hundred allosteric entries available at ASD. Among the multiple allosteric mechanisms known and the different effectors (other proteins, small-molecules, phosphorylation, etc), we chose to focus on small-molecule ligands, as these are the best candidates to be mimicked by therapeutic drugs [4,14]. Moreover, the approach presented here is based on the idea that changes in protein flexibility upon ligand binding can be related to allosteric and regulatory effects [1,24,36-38]. A simple computational way to estimate protein structural flexibility is the use of Normal Mode Analysis (NMA) [32,35,38,39]. In this case, however, we were not interested in measuring absolute flexibility values but the change in flexibility that occurs when a ligand binds to the protein structure in a particular location, in a similar fashion to the approach developed by Ming and Wall [32]. Once we had gathered and filtered allosteric proteins of known structure, we tested if changes in flexibility could be linked to the presence of the allosteric ligand. Experiments were performed using different molecular representations of the small-molecule ligands, and across different ranges of normal modes. Moreover, as a control we simulated the presence of ligands in alternative binding sites. This helped in parameterizing the methodology and made it applicable to cases where there is no *a priori* knowledge on the allostery the protein may present. Besides evaluating the overall results on a set of allosteric proteins, we took a closer look into particularly interesting cases.

Results and discussion

Gathering structural data on allosteric sites

To study allosteric sites from a structural perspective we first gathered the available data. We started by integrating the 146 allosteric site entries that were, at the time of this writing, annotated in the AlloSteric Database (ASD) [13] with another 72 allostery examples we had previously found in the literature. We proceeded to filter and cluster the data set as described in the Methods section to avoid overrepresentation [40] and low quality structures, turning the inital 213 cases into a total of 91 representative proteins where both the structure and location of the allosteric ligand are known.

Allosteric-ligand presence affects protein flexibility

Our first experiment aimed at quantifying the number of proteins in our data set that undergo a significant change in flexibility when the allosteric ligand is bound. However, known allostery cases show large diversity in their mechanisms [12] and we did not expect a positive result on the complete data set, since in many cases the allosteric effect may not be primarily driven by changes in local or overall flexibility but specific conformational changes, oligomerization or other mechanisms may be more relevant [41].

As explained in the Background section, we have chosen to estimate flexibility using Normal Mode Analysis (NMA). When applying NMA, calculated low-frequency modes reflect large collective oscillations of the protein structure and high-frequency modes reflect small local fluctuations [39]. Even though for most cases it has been shown that low-frequency normal modes are better descriptors of allosteric effects [42], we made no *a priori* assumption on the set of normal modes that would be more appropriate to detect an allosteric effect upon ligand binding for the ample protein set studied here. Thus, we decided to explore this parameter by using different ranges of normal modes, as described in the Methods section.

We used the calculated normal modes to predict B-factors [39], as this is a standard quantity for the estimation of protein flexibility [38]. Briefly, NMA calculations were performed for proteins in our data set both in the presence and absence of the allosteric ligand. For each protein, Cα B-factors derived from both conditions were compared and considered to be significantly different if the Wilcoxon-Mann-Whitney test returned a $p\text{-}value < 0.05$.

The results are displayed in Figure 1 and show that for the majority of the data set protein flexibility is significantly affected by the presence of the allosteric ligand. For most cases the effect was only observed when low-frequency normal modes were considered, as expected [35]. However, there are exceptions like the ribonucleotide reductase from *Thermotoga maritima* ([PDB:1XJF]), for which the allosteric effect has been described to be related to the local stabilization of three loops in the structure [4,17] and was captured only by high-frequency normal modes in our calculations.

Effect of ligand representation on the NMA results

We performed a second experiment to measure how different the results from this approach would be if we used a simplified molecular representation instead of the full-atom ligand, given the fact that knowledge on the ligand structure may not always be available. Moreover, a predictive approach that does not require information on the ligand molecule has a much larger field of application (*e.g.* structural genomics) paving the way for the discovery

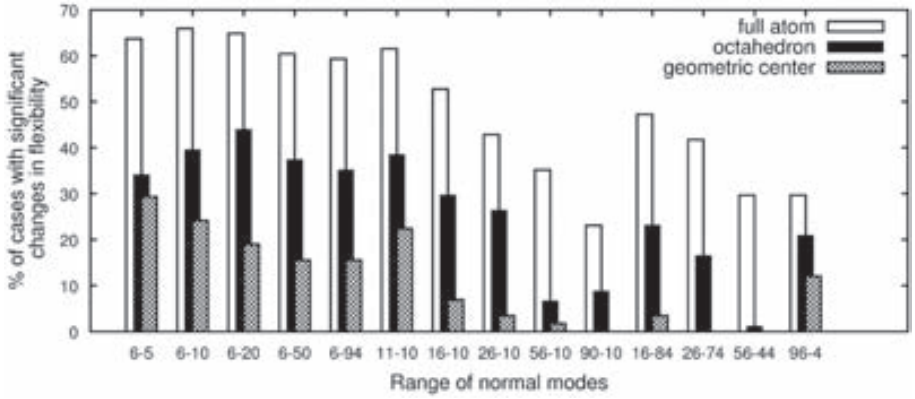

Figure 1 Ligand simulation. Identification of changes in flexibility upon ligand binding when using different ligand representations. Normal-mode range X-Y corresponds to the initial X modes being skipped and the next Y modes being taken into account.

of novel and pharmacologically interesting allosteric sites. Another interesting possibility that would open up is the detection of serendipitous allosteric sites, which despite having no natural ligand effectively become an allosteric site given the presence of an appropriate 'opportunistic' ligand [1].

We tested two representations of ligands: a single dummy atom located at the ligand geometric center and a set of 6 dummy atoms located at the vertices of an octahedron around the geometric center, as explained in the Methods section.

Figure 1 shows that for most cases the single dummy atom at the ligand geometric center is not able to trigger a significant change in flexibility during the simulations, while the octahedron exerts an effect much closer to that of the full-atom ligand. From a methodological point of view, simulating ligands in a simplified form allowed us to perform control experiments which are described below.

Predicting ligand-binding pockets and selection of normal mode range

To further develop a predictive approach, we used the LIGSITE^{cs} program [43] to predict the putative ligand-binding sites on the protein structure. Different programs are available for this task with very good performance in general as shown in recent reviews [44-46]. We chose LIGSITE^{cs} because pockets are predicted based only on the shape of the protein surface; programs incorporating more parameters (*e.g.* evolutionary conservation, druggability) could have improved but also biased our results.

We predicted the location of up to 8 ligand-binding pockets per protein and performed NMA to check if any of the predicted pockets had a significant effect upon protein flexibility when occupied by a small-molecule ligand, as described in the Methods section.

A pocket that presents no ligand (*i.e.* appears empty in the original structure) may nevertheless display a significant change in overall flexibility if occupied by a ligand representation when performing the normal-modes calculation. It would be wrong to consider this directly an error, since native ligands may exist that bind this pocket even if there is none present in the particular experimental structure under study. A few examples are mentioned in the next section, where we found pockets that affect protein flexibility and, although they are indeed not allosteric sites, they are active sites or other biologically relevant sites. Nevertheless, to guide the definition of the model we needed an error propensity estimate for the different parameters tested, i.e. range of normal modes and ligand molecular representation. If all pockets predicted on the protein surface would be found to affect significantly the flexibility, the corresponding parameters would be rendering the method too sensitive (low specificity) and prone to present false positives. Based on this argument, we estimated error propensity (*ep*) for each range of normal modes and the two ligand representations using the following *ad hoc* equation:

$$ep = \frac{p_7 + p_8 + 1}{p_1 + p_2 + 1} \quad (1)$$

where p_x is the number of cases in which x pockets were predicted to be significantly affecting the overall protein flexibility. Note that this equation does not formally stand for an error but simply gives an idea of the likelihood of having false positives. The results are displayed in Figure 2 and show that the octahedron representation, combined with the lowest frequency normal modes, leads to a higher specificity (lower number of pockets significantly affecting overall dynamics) than the single dummy atom at the geometric center. We then decided to continue our work using the octahedron representation of ligands together with normal modes in the range 6-20.

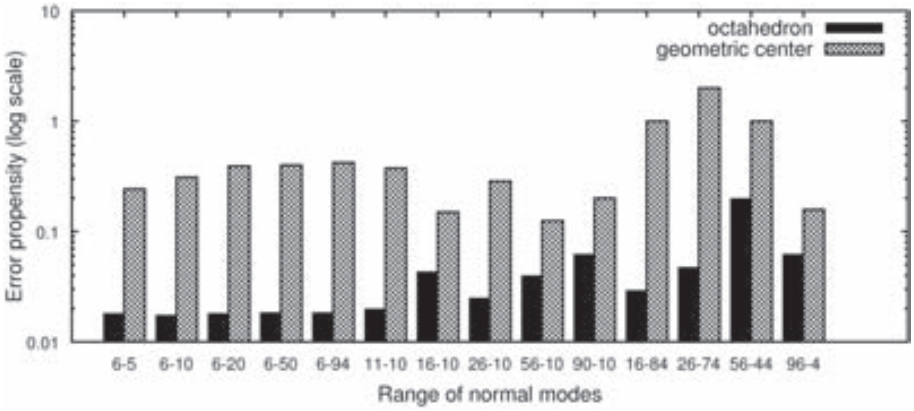

Figure 2 Error propensity estimation. Ratios between the number of cases with 1 or 2 significant pockets and the number of cases with 7 or 8 significant pockets. Normal-mode range X-Y corresponds to the initial X modes being skipped and the next Y modes being taken into account.

Overall performance when predicting allosteric sites

At the time of this writing, no large-scale study attempting the prediction of the allosteric-site location in known allosteric proteins has been published. Recent work by Demerdash and coworkers aimed at predicting the residues involved in the propagation of allosteric signals within a protein structure for a set of 16 different proteins [26]. Quite distinctly, our method follows a drug-discovery oriented approach where the intention is to pinpoint specific protein pockets that present a high potential for affecting biological function. From that perspective our method is comparable to the one developed by Ming and Wall for finding functional sites, as it exploits NMA to assess the differences in flexibility between ligand-bound and unbound states of a protein [33,47]. However, their approach differs from ours in multiple major points, including the sampling of protein sites, the parametrization of probes and of their interaction with the protein and the approach by which the perturbation of protein dynamics is assessed. The method described in this paper is also similar to the approach recently published by Mitternacht and coworkers, where they measured the 'binding leverage', or ability of a binding site to couple to the intrinsic motions of a protein, on a set of 15 allosteric proteins [35].

In this context, any pocket with a biological regulatory role would be suited for the analysis, but we chose to focus on allosteric sites since these are possibly the most interesting, albeit complex, regulatory sites to approach. Starting from a data set containing 91 proteins, we measured the rate of success of our approach to identify allosteric sites as follows. First, we discarded a total of 33 proteins for which no single LIGSITE^{cs} predicted pocket matched the allosteric site, leaving a total of 58 cases to work with (63,7%). The rational for discarding these proteins is that the present analysis is not concerned with the ability of a specific program to detect a cavity but with the ability of our approach to identify the cavity, among those detected, that corresponds to the allosteric site. Indeed, it has been previously observed that not all allosteric sites are predicted to be potential ligand-binding cavities by common algorithms [35]. There can be different reasons for this, for example the allosteric site may be deeply buried in the protein, may display a planar shape or be located at the interface of subunits, making it difficult for the pocket-prediction algorithm to detect its presence.

A total of 464 pockets were predicted on the surface of the 58 proteins (8 per protein). The chance of randomly selecting an allosteric site is low, given that only 13% of these pockets (one per protein) matched the location of an allosteric site (*i.e.* its center less than 5 Å away from the allosteric ligand; if more than one pocket matched the ligand position within this cut-off, the closest was chosen). After performing the analysis of normal modes, 117 pockets display a significant effect on the overall protein flexilibity upon ligand binding (set F in Table 1). The chance of success (positive predictive value) more than doubles with the incorporation of this analysis, with 27% of these 117 pockets matching an allosteric site. Furthermore, we integrated these results with our previous work on protein-pocket conservation by selecting pockets that display at least 50% structural conservation, as defined previously [23]. Interestingly, considering protein conservation alone (set S in Table 1) results in a slightly lower positive predictive value than considering only flexibility. While the two measures show the same specificity, using the effect on flexibility as criterion leads to a slightly higher sensitivity than using the structural-conservation feature. The double-filtered set, combining the effect on flexibility with a high structural conservation (set FS), contains only 36 pockets, of which 15 (42%) match an allosteric site (Table 1). This represents a nearly four times larger positive predictive value than 'random' selection

Table 1 Prediction results on protein allosteric sites

Set	TP+FP	TP	FP	FN	Sensitivity	Specificity	Accuracy	PPV
Total	464	58	406	0	1.00	0.00	0.13	0.13
F	117	32	85	26	0.55	0.79	0.76	0.27
S	108	24	84	34	0.41	0.79	0.75	0.22
FS	36	15	21	43	0.26	0.95	0.86	0.42
c123	174	44	130	14	0.76	0.68	0.69	0.25
c123F	74	29	45	29	0.50	0.89	0.84	0.39
c123S	55	22	33	36	0.38	0.92	0.85	0.40
c123FS	30	14	16	44	0.24	0.96	0.87	0.47
c1	58	26	32	32	0.45	0.92	0.86	0.45
c1F	42	22	20	36	0.38	0.95	0.88	0.52
c1S	25	15	10	43	0.26	0.98	0.89	0.60
c1FS	20	13	7	45	0.22	0.98	0.89	0.65

Results on this table refer to the subset of 58 proteins for which LIGSITE^{cs} was able to predict a ligand-binding pocket in the position of the allosteric site. TP: true positive; TN: true negative; FP: false positive; FN: false negative; PPV: positive predictive value. Sensitivity: TP/(TP+FN); specificity: TN/(TN+FP); accuracy: (TP+TN)/(TP+FN+TN+FP); PPV or precision: TP/(TP+FP). The total number of pockets considered, predicted by LIGSITE^{cs}, is 464 (8 per protein). F corresponds to sets including a change in flexibility as selection criterion; S corresponds to sets including high structural conservation as selection criterion; c123 refers to sets considering only the three largest pockets predicted by LIGSITE^{cs}; c1 refers to sets considering only the largest predicted pocket.

within the 464 identified cavities, at the expense of reducing further the sensitivity of the approach, i.e. decreasing drastically the number of false positives but increasing also the number of false negatives.

However, it will not be common to select up to eight pockets per protein as potential allosteric sites. A researcher working on a particular protein without *a priori* knowledge on its regulatory mechanism will probably keep the first three largest pockets predicted by default [43] or, as Thornton and coworkers explain for the case of active sites [21], the largest pocket will usually be the best bet. In those two scenarios, the tendency shown for the complete set of predicted pockets is conserved (Table 1). When keeping the first 3 pockets (set c123), the chance to match an allosteric site (positive predictive value) goes from 25% to 39% when using the flexibility criterion and up to 47% when incorporating structural conservation as well. Out of the 58 allosteric sites, however, 14 are not found within the c123 set. Likewise, when selecting only the first and largest pocket, the inital success rate goes from 45% to 52% when considering the effect on flexibility upon binding (set c1F) and to 65% when structural conservation is also required. Note that between sets FS and c1FS the number of false positives decreases by three-fold, while only two additional false negatives are added.

We considered only allosteric sites as desirable matches. However, other pockets with biological functions were matched by our criteria, as described further below on a few particular examples. The performance of this approach might be improvable using other pocket prediction programs or a combination of them. However, performance of pocket prediction methods does not vary largely, as shown by a recent large-scale comparison [45]. Our study represents the largest test to date (58 non-redundant proteins in complex with their corresponding small-molecule allosteric ligands) proving the concept that changes in overall flexibility upon ligand binding are relevant identifiers for some allosteric sites, and these effects can be captured in many cases with the simple approach described here. In addition, we further show (see also [23]) that evaluation of the structural conservation of the candidate pockets may contribute as much to the identification of the allosteric site.

Biological examples

As mentioned in the Background section, allostery can work through many different mechanisms. Thus, we consider it important, besides the overall results presented above, to explain the results for a few proteins in more detail. The following section should help to better illustrate the relevance of incorporating a flexibility measure when studying allosteric systems and predicting the location of allosteric sites.

Glyceraldehyde 3-phosphate dehydrogenase

Aldehyde dehydrogenases (ALDH) are found across all kingdoms of life. They play a vital role in multiple cellular processes, including glycolysis, detoxification and embryogenic development. A distinct family within the ALDH superfamily consists of the non-phosphorylating glyceraldehyde-3-phosphate dehydrogenases (GAPN), which catalyze the phosphate-independent irreversible oxidation of glyceraldehyde 3-phosphate (GAP) to 3-phosphoglycerate using NAD(P) as a cosubstrate. Unlike

other proteins in the GAPN family, the enzyme of the hyperthermophilic Archaeum *Thermoproteus tenax* (Tt-GAPN) is regulated by a set of inhibitors (NADH, NADP(H) and ATP) and activators (AMP, ADP, glucose 1-phosphate and fructose 6-phosphate (F6P)) which decrease or increase, respectively, the affinity for NAD. This suggests that Tt-GAPN plays a crucial role in regulating the carbohydrate catabolism in *T. tenax* [48].

All different activators bind to the same allosteric site, which is located more than 20 Å away from both the active site and the cosubstrate-binding site of any monomer of the tetramer [49]. The activator binding site is located at the interface between the tetramerization domain and the cosubstrate binding/catalytic domains. It is also observed that the allosteric ligands are in direct contact with 3 or even four monomers in the protein complex, indicating a role in the stabilization of the complex. This role probably combines with the detected effect on flexibility to influence enzyme kinetics, as no large conformational change is observed when comparing the ligand-bound and ligand-free structures besides a rearrangement of the tetramerization domain with respect to the cosubstrate binding/catalytic domain [49].

In our analysis, PKT5 (the fifth largest pocket predicted) matched the location of the allosteric effector F6P, as shown in Figure 3. When a ligand was simulated occupying this pocket, using the octahedron representation, the overall flexibility of the protein was significantly affected on all ranges including the lowest frequency modes (*p-value* <= 0.001). No other pocket presented the same behaviour (Figure 3), not even the largest pocket (PKT1), which matches the position of cofactor NADP. Given the 'hinge-like' position of the activator binding site and the variety of ligands that it can accomodate, we consider this case a good example to speculate that the actual position of the ligand in the structure plays a major role in its effect on the protein activity, beyond the particular chemical properties of the ligand itself that may be important for binding.

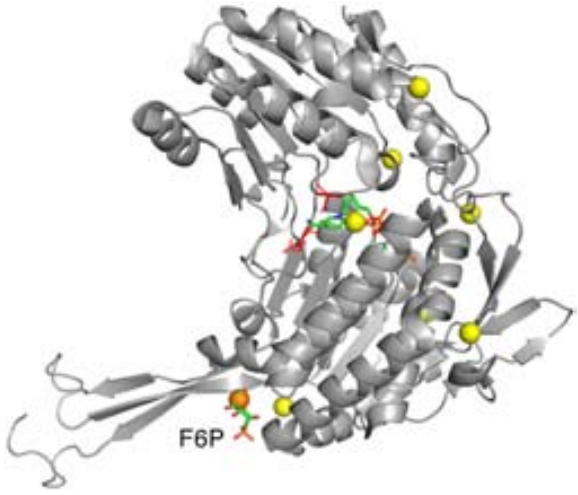

Figure 3 Glyceraldehyde 3-phosphate dehydrogenase. Predicted pockets and ligands on *Thermoproteus tenax* glyceraldehyde 3-phosphate dehydrogenase (TtGAPN). Only a single protein monomer is shown ([PDB:1UXR]). NADPH at the active site and activator fructose-6-phosphate (F6P) at the allosteric site are shown in 'sticks' representation, while residues in red correspond to the CSA [21] active site annotations. Predicted pockets (geometric centers) are shown in 'spheres' representation, the pocket in orange color affected protein flexibility significantly according to our simulations, while yellow did not.

PDK1 kinase

PDK1 kinase is a key regulator of AGC kinases, which play crucial roles in physiological processes relevant to metabolism, growth, proliferation and survival [50]. This protein is regulated allosterically by the binding of a phosphopeptide which Biondi and coworkers managed to mimic with a low-molecular-weight activator [51] and further solved the structure of the complex [52]. As shown in Figure 4, the largest pocket (PKT1) matches the binding site of ATP. In our analysis, PKT1 affected protein flexibility significantly on most normal-mode ranges. The second largest pocket predicted (PKT2) matches the location of the allosteric activator (PS48) at the HM/PIF binding site. According to the analysis, based on the lowest-frequency normal modes (6-5 range) PKT2 significantly affects overall protein flexibility if occupied by a ligand.

Another predicted pocket (PKT7), appears to significantly affect protein flexibility on most normal-mode ranges when occupied during the NMA. This pocket

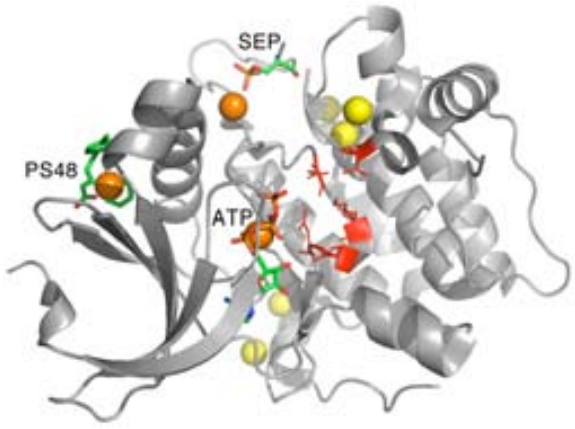

Figure 4 PDK1 kinase. Predicted pockets and ligands on PDK1 kinase ([PDB:3HRF]). Ligands (ATP, allosteric effector PS48) and the modified residue phosphoserine (SEP) are shown in 'sticks' representation, residues in red correspond to the CSA [21] active site annotations. Predicted pockets (geometric centers) are shown in 'spheres' representation, orange pockets affected protein flexibility significantly according to our simulations, but yellow did not.

is not occupied by any ligand in the original structure ([PDB:3HRF]), but it does match the position of a phosphoserine (SEP) in the activation loop of PDK1, as shown in Figure 4. Also in this pocket, residue THR226 is considered a crucial element of the allosteric mechanism of this protein, given that mutation of this residue inhibits activation without inhibiting binding [52]. These results indicate that stabilization of this protein region would have an effect on the overall flexibility of the protein, linking it to a regulatory function which correlates with what has been observed previously based on deuterium exchange and other experimental procedures [52]. The other 5 pockets predicted on this structure were not found to significantly affect protein flexibility.

HIV reverse transcriptase

Non-nucleoside reverse transcriptase inhibitors (NNRTIs) are key elements of the so-called HAART (Highly Active Antiretroviral Therapy) multi-drug treatments against HIV-1 infection. However, rapid mutation of HIV-1 compromises the efficacy and durability of HAART. This high mutation rate fuels the need to discover novel agents with better activity profiles against HIV-1 reverse transcriptase (RT) and its most common mutants. In this context, Anthony and coworkers have developed substituted tetrahydroquinolines which are potent allosteric inhibitors of HIV-1 RT and some of its key mutants [53].

In our normal-mode analysis the two significant pockets matched the position of the allosteric ligands, as displayed in Figure 5. The allosteric site is located in a 'hinge-like' position between domains, a position which has even been exploited for the engineering of regulatory sites as well [54]. A ligand bound in this position would easily perturb the low frequency modes of vibration of the protein, thus affecting its overall flexibility and subsequently altering protein function. All other pockets predicted on this structure were found not to affect protein flexibility significantly, meaning that a hypothetical blind drug-design approach focused on the significant pockets from the NMA would have been successful. This is an excellent example showing that the combination of pocket prediction and NMA may pinpoint the location of the allosteric/regulatory site based solely on structural data.

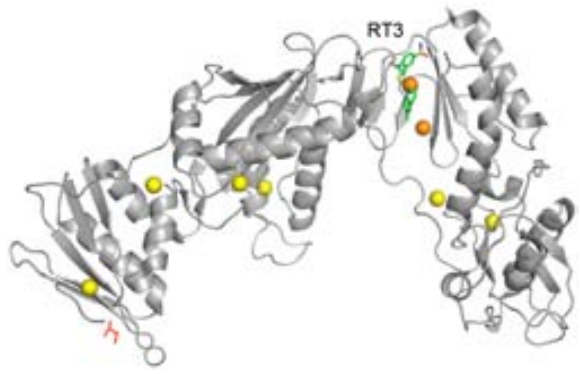

Figure 5 HIV reverse transcriptase. Predicted pockets and ligands on HIV reverse transcriptase ([PDB:3I0R]). Compound RT3 at the allosteric site is shown in 'sticks' representation, while residues in red correspond to the CSA [21] active site annotations. Predicted pockets (geometric centers) are shown in 'spheres' representation, the pocket in orange color affected protein flexibility significantly according to our simulations, while yellow did not.

L-lactate dehydrogenase

When glycolysis takes place under anaerobic conditions, pyruvate is reduced to L-lactate, a reaction that is catalyzed by L-lactate dehydrogenase (LDH). In contrast to their mammalian counterparts, some bacterial LDHs display allosteric regulation by fructose 1,6-bisphosphate (FBP) [55]. Iwata and co-workers solved the structure of *Bifidobacterium longum* LDH in both active (R) and inactive (T) states, co-crystallized with the allosteric activator [56]. A significant difference can be observed between the B-factors of both structures, suggesting an overall change in flexibility being part of the allosteric mechanism.

We mentioned this protein in our previous work [23], where we found the allosteric site to be structurally conserved although no signal of sequence conservation was found. In the current analysis, the only pocket that perturbed the overall flexibility of LDH when we simulated the presence of a ligand was the second largest pocket (PKT2), which is also the one closest to the allosteric site, as displayed in Figure 6. We did not consider this case as a 'match' in the large-scale results shown in Table 1 because the pocket geometric center is 6.6 Å away from the allosteric ligand, thus failing the pre-defined threshold of 5 Å. However, after visual inspection we considered this case relevant because the ligand is occupying the same large pocket, even if it is not located precisely at the pocket center defined by LIGSITE^{cs}.

No other pocket on this protein displayed an effect related to flexibility according to our calculations when considering the normal mode range 6-20, not even those pockets matching the location of the active site or other ligands.

Given that animal LDHs are not regulated allosterically, this protein/pocket could be an excellent target for antimicrobial compounds. To further explore this idea, we analyzed the human LDH homolog ([PDB:1I10]) as well, which shows a sequence identity of 37.7% and a local RMSD of 1.04 Å according to the SUPERPOSE web server [57] when compared to *Bifidobacterium longum* LDH. On the human protein, which is not regulated allosterically, the pocket equivalent to the allosteric site in the bacterial homolog did not produce a significant effect on

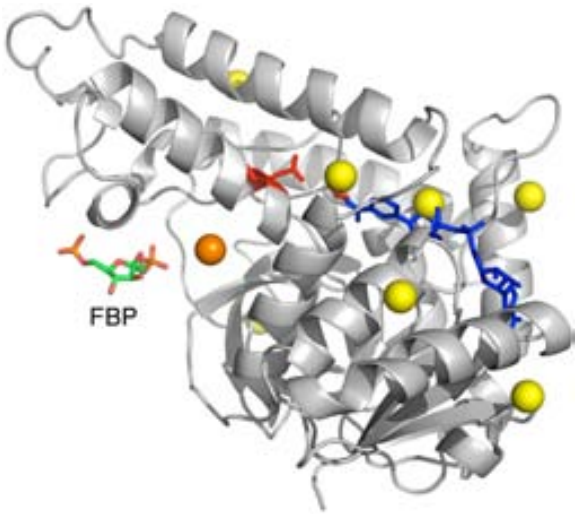

Figure 6 L-lactate dehydrogenase. Predicted pockets and ligands on L-lactate dehydrogenase ([PDB:1LTH]). Fructose-1,6-bisphosphate (FBP) at the allosteric site is shown in 'sticks' representation as well as NAD which is colored blue, while residues in red correspond to the CSA [21] active site annotations. Predicted pockets (geometric centers) are shown in 'spheres' representation, the pocket in orange color affected protein flexibility significantly according to our simulations, while yellow did not.

flexibility according to our calculations. It is remarkable that a coarse approximation such as this (based on Cαs and NMA) is able to distinguish that the presence of the allosteric ligand has a significant effect on the bacterial protein flexibility but not on its human homolog.

Conclusions

In this article we have proposed a very simple approach exploiting changes in protein flexibility upon ligand binding to predict the presence and location of allosteric sites. We tested the methodology on a non-redundant set of 58 proteins achieving in the best case a success rate (positive predictive value) of 65%, with a sensitivity of 0.22. Furthermore, we analyzed four cases in more detail, revealing how the coarse-grained approach described here is able to capture the effect triggered by the allosteric ligand, matching the current literature. The structural analysis proposed here could help medicinal chemists and other researchers on their way through the promising field of allosteric-drug design.

Methods

To ensure that the quality and nature of the selected structural data was appropriate to our study, we discarded structures with a resolution lower than 3 Å or with a G-Factor lower than -1, as calculated by PROCHECK [58]. We conservatively defined a non-redundant data set to avoid possible bias in the results that may arise from overrepresentation of any protein family [40]. Clustering was performed with the BLASTCLUST program [59] using a threshold of 30% sequence identity, which grouped the 213 initial entries into 91 groups. We then selected the highest resolution structure of each group as its representative and defined a non-redundant data set which contained a total of 91 distinct allosteric proteins, for which the structure and location of both the allosteric site and ligand were known.

Normal mode analysis (NMA) was performed on the protein crystallographic structures with and without a probe ligand, in its three different representations (full atom, octahedron and geometric center, see below). The simplified ligand representations (octahedron and geometric center) were alternatively placed in each of the eight predicted pockets. The NMA was based on the implementation of Sanejouand and coworkers [39,60] using the programs PDBMAT and DIAGRTB. The calculation involves the diagonalization of the mass-weighted Hessian (**H**) of the potential energy function V. Following Tirion's Elastic Network Model [61], the potential energy V is simply described as a set of harmonic springs of equal strength k linking every pair of Cα atoms with a distance smaller than R_c in the crystallographic structure:

$$V = \sum_{\substack{r_{ij}^0 < R_c \\ i<j}} k(r_{ij} - r_{ij}^0)^2 \quad (2)$$

where r_{ij}^0 is the Euclidean distance between atoms i and j in the crystallographic structure and R_c and k were given in this study the values 10 Å and 1 Kcal mol^{-1} Å^{-2}, respectively. Note that this energy function was designed in such a way that it does not require energy minimization of the X-ray structure prior to the normal-mode calculation since the X-ray structure is the minimum of the function. Although this method uses very gross approximations (reduction to Cα atoms, extremely simple energy function, no solvent), it has proven to perform surprisingly well in front of both more complex approximations and experimental data (B-factors) [39,60,62].

The eigenvectors and eigenvalues of **H** correspond to the normal modes, characterizing the direction and amplitude of the vibrational motion, and frequencies of vibration, respectively. They can be used to calculate mean-square displacements of the atomic cartesian coordinates x ($\langle x^2 \rangle$) as:

$$\langle x_i^2 \rangle = \frac{K_B T}{m_i} \sum_{j=1}^{n_v} \frac{a_{ij}^2}{w_j^2} \quad (3)$$

where x_i is the coordinate i, m_i the corresponding mass, K_B Boltzmann's constant, T the temperature, n_v the number of modes considered, $w_j/2\pi$ the frequency of normal mode j and a_{ij} the coordinate i of normal mode j. The

resulting mean-square radial displacements of atom positions $\langle r^2 \rangle$ can in turn be used to estimate atomic B-factors as:

$$B = \frac{8\pi^2}{3} \langle r^2 \rangle \quad (4)$$

Low-frequency modes reflect large collective or delocalized motions in the protein structure, while high-frequency modes reflect small vibrations in localized regions. We estimated B-factors for the ligand-bound and unbound protein structures using different ranges of normal modes to explore this variable. Ranges were named using two numbers X-Y: Starting from the low frequency modes, X is the number of modes that are skipped and Y is the number of normal modes that are taken into account. The first six normal modes, with zero frequency, are of no interest as they represent rigid-body translation and rotation. The ranges tested were: 6-5, 6-10, 6-20, 6-50, 6-94, 11-10, 16-10, 26-10, 56-10, 90-10, 16-84, 26-74, 56-44 and 96-4.

We prepared protein coordinate files for NMA as follows: (1) Protein chains in direct contact with the allosteric ligand (*i.e.* multiple residues within 3.0 Å) were selected and atoms belonging to other chains or molecules in the structure were removed. (2) The LIGSITE^{cs} program [43] was used to predict up to 8 pockets per structure. (3) After pocket prediction, protein structures were parsed to keep Cα atoms only.

We took the first 100 normal modes for each protein and ligand representation: **apo**, the Cα only 'apo' protein crystallographic structure (allosteric ligand is not present); **ligand**, same protein structure as in 'apo' but including the allosteric ligand (or a simplified molecular representation) in the allosteric site; **PKTX**, same protein structure as in 'apo' plus a simplified molecular representation of a ligand occupying the predicted pocket number X (1 to 8).

In the last case, the molecular representation of the ligand was located at the pocket geometric center, as predicted by LIGSITE^{cs}.

The ligand molecule during NMA was simulated in different ways: **full atom**, all atoms in the ligand molecule are included in the calculation; **geometric center**, a single dummy atom located at the ligand-pocket geometric center is considered; **octahedron**, the ligand's presence is simulated by a dummy atom positioned at the geometric center and six extra dummy atoms located at 4 Å distance from the center on both sides of each axis (*i.e.* forming the vertices of a regular octahedron).

For each protein-ligand pair, calculated B-factors for the Cα protein atoms of the **apo** structure were compared to those obtained for the same atoms in the configurations including real or simulated ligands to test for significant changes in flexibility using the Wilcoxon-Mann-Whitney test. Differences with a *p-value* < 0.05 were considered significant.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AP and XD conceived the study and wrote the manuscript. AP carried out the computational work. Both authors read and approved the final manuscript.

Acknowledgements
The authors dedicate this paper to Professor Wilfred F. van Gunsteren on the occasion of his 65th birthday. This project is supported by funding under the Seventh Research Framework Programme of the European Union (ref. HEALTH-F3-2009-223101). AP acknowledges the FPU Scholarship from MICINN, Spanish Gov.

Received: 26 April 2012 Accepted: 17 October 2012
Published: 25 October 2012

References

1. Hardy JA, Wells JA: **Searching for new allosteric sites in enzymes.** *Curr Opin Struct Biol* 2004, **14**(6):706–715. [http://dx.doi.org/10.1016/j.sbi.2004.10.009].
2. Monod J: *Chance and Necessity: Essay on the Natural Philosophy of Modern Biology*. New York: Vintage Books; 1977. ISBN 978-0394718255.
3. Goodey NM, Benkovic SJ: **Allosteric regulation and catalysis emerge via a common route.** *Nat Chem Biol* 2008, **4**(8):474–482. [http://dx.doi.org/10.1038/nchembio.98].
4. Peracchi A, Mozzarelli A: **Exploring and exploiting allostery: Models, evolution, and drug targeting.** *Biochim Biophys Acta* 2011, **1814**(8):922–933. [http://dx.doi.org/10.1016/j.bbapap.2010.10.008].
5. Möhler H, Fritschy JM, Rudolph U: **A new benzodiazepine pharmacology.** *J Pharmacol Exp Ther* 2002, **300:**2–8.
6. Raddatz R, Schaffhauser H, Marino MJ: **Allosteric approaches to the targeting of G-protein-coupled receptors for novel drug discovery: a critical assessment.** *Biochem Pharmacol* 2007, **74**(3):383–391. [http://dx.doi.org/10.1016/j.bcp.2007.05.007].
7. Nussinov R, Tsai CJ, Csermely P: **Allo-network drugs: harnessing allostery in cellular networks.** *Trends Pharmacol Sci* 2011, **0:**0. [http://dx.doi.org/10.1016/j.tips.2011.08.004].
8. Grimsby J, Sarabu R, Corbett WL, Haynes NE, Bizzarro FT, Coffey JW, Guertin KR, Hilliard DW, Kester RF, Mahaney PE, Marcus L, Qi L, Spence CL, Tengi J, Magnuson MA, Chu CA, Dvorozniak MT, Matschinsky FM, Grippo JF: **Allosteric activators of glucokinase: potential role in diabetes therapy.** *Science* 2003, **301**(5631):370–373. [http://dx.doi.org/10.1126/science.1084073].
9. Doliba NM, Qin W, Najafi H, Liu C, Buettger CW, Sotiris J, Weik-Collins H, Li C, Stanley CA, Wilson DF, Grimsby J, Sarabu R, Naji A, Matschinsky FM: **Glucokinase Activation Repairs Defective Bioenergetics of Islets of Langerhans isolated from Type-2-Diabetics.** *Am J Physiol Endocrinol Metab* 2011, **0:**0. [http://dx.doi.org/10.1152/ajpendo.00218.2011].
10. Kohl A, Amstutz P, Parizek P, Binz HK, Briand C, Capitani G, Forrer P, Plückthun A, Grütter: **Allosteric inhibition of aminoglycoside phosphotransferase by a designed ankyrin repeat protein.** *Structure* 2005, **13**(8):1131–1141. [http://dx.doi.org/10.1016/j.str.2005.04.020].
11. Wishart DS, Knox C, Guo AC, 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**(Database issue):D901—D906. [http://dx.doi.org/10.1093/nar/gkm958].
12. Laskowski RA, Gerick F, Thornton JM: **The structural basis of allosteric regulation in proteins.** *FEBS Lett* 2009, **583**(11):1692–1698. [http://dx.doi.org/10.1016/j.febslet.2009.03.019].
13. Huang Z, Zhu L, Cao Y, Wu G, Liu X, Chen Y, Wang Q, Shi T, Zhao Y, Wang Y, Li W, Li Y, Chen H, Chen G, Zhang J: **ASD: a comprehensive database of allosteric proteins and modulators.** *Nucleic Acids Res* 2011, **39**(Database issue):D663–D669. [http://dx.doi.org/10.1093/nar/gkq1022].
14. Zorn JA, Wells JA: **Turning enzymes ON with small molecules.** *Nat Chem Biol* 2010, **6**(3):179–188. [http://dx.doi.org/10.1038/nchembio.318].
15. Heredia VV, Thomson J, Nettleton D, Sun S: **Glucose-induced conformational changes in glucokinase mediate allosteric**

regulation: transient kinetic analysis. *Biochemistry* 2006, **45**(24):7553–7562. [http://dx.doi.org/10.1021/bi060253q].
16. Maita N, Okada K, Hatakeyama K, Hakoshima T: **Crystal structure of the stimulatory complex of GTP cyclohydrolase I and its feedback regulatory protein GFRP.** *Proc Natl Acad Sci U S A* 2002, **99**(3):1212–1217. [http://dx.doi.org/10.1073/pnas.022646999].
17. Reichard P: **Ribonucleotide reductases: substrate specificity by allostery.** *Biochem Biophys Res Commun* 2010, **396:**19–23. [http://dx.doi.org/10.1016/j.bbrc.2010.02.108].
18. Cooper A, Dryden DT: **Allostery without conformational change. A plausible model.** *Eur Biophys J* 1984, **11**(2):103–109.
19. Tsai CJ, del Sol, A, Nussinov R: **Allostery: absence of a change in shape does not imply that allostery is not at play.** *J Mol Biol* 2008, **378:**1–11. [http://dx.doi.org/10.1016/j.jmb.2008.02.034].
20. Gunasekaran K, Ma B, Nussinov R: **Is allostery an intrinsic property of all dynamic proteins?** *Proteins* 2004, **57**(3):433–443. [http://dx.doi.org/10.1002/prot.20232].
21. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32**(Database issue):D129–D133. [http://dx.doi.org/10.1093/nar/gkh028].
22. Mistry J, Bateman A, Finn RD: **Predicting active site residue annotations in the Pfam database.** *BMC Bioinformatics* 2007, **8:**298. [http://dx.doi.org/10.1186/1471-2105-8-298].
23. 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. [http://dx.doi.org/10.1186/1472-6807-10-9].
24. Daily MD, Gray JJ: **Local motions in a benchmark of allosteric proteins.** *Proteins* 2007, **67**(2):385–399. [http://dx.doi.org/10.1002/prot.21300].
25. Daily MD, Gray JJ: **Allosteric communication occurs via networks of tertiary and quaternary motions in proteins.** *PLoS Comput Biol* 2009, **5**(2):e1000293. [http://dx.doi.org/10.1371/journal.pcbi.1000293].
26. Demerdash ONA, Daily MD, Mitchell JC: **Structure-based predictive models for allosteric hot spots.** *PLoS Comput Biol* 2009, **5**(10):e1000531. [http://dx.doi.org/10.1371/journal.pcbi.1000531].
27. del Sol A, Tsai CJ, Ma B, Nussinov R: **The origin of allosteric functional modulation: multiple pre-existing pathways.** *Structure* 2009, **17**(8):1042–1050. [http://dx.doi.org/10.1016/j.str.2009.06.008].
28. Freire E: **Can allosteric regulation be predicted from structure?.** *Proc Natl Acad Sci U S A* 2000, **97**(22):11680–11682. [http://dx.doi.org/10.1073/pnas.97.22.11680].
29. Liu J, Nussinov R: **Allosteric effects in the marginally stable von Hippel-Lindau tumor suppressor protein and allostery-based rescue mutant design.** *Proc Natl Acad Sci U S A* 2008, **105**(3):901–906. [http://dx.doi.org/10.1073/pnas.0707401105].
30. Kidd BA, Baker D, Thomas WE: **Computation of conformational coupling in allosteric proteins.** *PLoS Comput Biol* 2009, **5**(8):e1000484. [http://dx.doi.org/10.1371/journal.pcbi.1000484].
31. Rader AJ, Brown SM: **Correlating allostery with rigidity.** *Mol Biosyst* 2011, **7**(2):464–471. [http://dx.doi.org/10.1039/c0mb00054j].
32. Ming D, Wall ME: **Quantifying allosteric effects in proteins.** *Proteins* 2005, **59**(4):697–707. [http://dx.doi.org/10.1002/prot.20440].
33. Ming D, Wall ME: **Interactions in native binding sites cause a large change in protein dynamics.** *J Mol Biol* 2006, **358:**213–223. [http://dx.doi.org/10.1016/j.jmb.2006.01.097].
34. Mitternacht S, Berezovsky IN: **A geometry-based generic predictor for catalytic and allosteric sites.** *Protein Eng Des Sel* 2011, **24**(4):405–409. [http://dx.doi.org/10.1093/protein/gzq115].
35. Mitternacht S, Berezovsky IN: **Binding leverage as a molecular basis for allosteric regulation.** *PLoS Comput Biol* 2011, **7**(9):e1002148. [http://dx.doi.org/10.1371/journal.pcbi.1002148].
36. Teague SJ: **Implications of protein flexibility for drug discovery.** *Nat Rev Drug Discov* 2003, **2**(7):527–541. [http://dx.doi.org/10.1038/nrd1129].
37. Balabin IA, Yang W, Beratan DN: **Coarse-grained modeling of allosteric regulation in protein receptors.** *Proc Natl Acad Sci U S A* 2009, **106**(34):14253–14258. [http://dx.doi.org/10.1073/pnas.0901811106].
38. Dykeman EC, Twarock R: **All-atom normal-mode analysis reveals an RNA-induced allostery in a bacteriophage coat protein.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2010, **81**(3 Pt 1):031908.
39. Delarue M, Sanejouand YH: **Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model.** *J Mol Biol* 2002, **320**(5):1011–1024.
40. Xie L, Bourne PE: **Functional coverage of the human genome by existing structures, structural genomics targets, and homology models.** *PLoS Comput Biol* 2005, **1**(3):e31. [http://dx.doi.org/10.1371/journal.pcbi.0010031].
41. Tsai CJ, Sol AD, Nussinov R: **Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms.** *Mol Biosyst* 2009, **5**(3):207–216. [http://dx.doi.org/10.1039/b819720b].
42. Zheng W, Brooks BR, Thirumalai D: **Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations.** *Proc Natl Acad Sci U S A* 2006, **103**(20):7664–7669. [http://dx.doi.org/10.1073/pnas.0510426103].
43. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.** *BMC Struct Biol* 2006, **6:**19. [http://dx.doi.org/10.1186/1472-6807-6-19].
44. Huang B: **MetaPocket: a meta approach to improve protein ligand binding site prediction.** *OMICS* 2009, **13**(4):325–330. [http://dx.doi.org/10.1089/omi.2009.0045].
45. Schmidtke P, Souaille C, Estienne F, Baurin N, Kroemer RT: **Large-scale comparison of four binding site detection algorithms.** *J Chem Inf Model* 2010, **50**(12):2191–2200. [http://dx.doi.org/10.1021/ci1000289].
46. Chen K, Mizianty MJ, Gao J, Kurgan L: **A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds.** *Structure* 2011, **19**(5):613–621. [http://dx.doi.org/10.1016/j.str.2011.02.015].
47. Ming D, Wall ME: **Predicting binding sites by analyzing allosteric effects.** *Methods Mol Biol* 2012, **796:**423–436. [http://dx.doi.org/10.1007/978-1-61779-334-9_23].
48. Brunner NA, Brinkmann H, Siebers B, Hensel R: **NAD+-dependent glyceraldehyde-3-phosphate dehydrogenase from Thermoproteus tenax. The first identified archaeal member of the aldehyde dehydrogenase superfamily is a glycolytic enzyme with unusual regulatory properties.** *J Biol Chem* 1998, **273**(11):6149–6156.
49. Lorentzen E, Hensel R, Knura T, Ahmed H, Pohl E: **Structural Basis of allosteric regulation and substrate specificity of the non-phosphorylating glyceraldehyde 3-Phosphate dehydrogenase from Thermoproteus tenax.** *J Mol Biol* 2004, **341**(3):815–828. [http://dx.doi.org/10.1016/j.jmb.2004.05.032].
50. Mora A, Komander D, van Aalten, D M F, Alessi DR: **PDK1, the master regulator of AGC kinase signal transduction.** *Semin Cell Dev Biol* 2004, **15**(2):161–170.
51. Engel M, Hindie V, Lopez-Garcia LA, Stroba A, Schaeffer F, Adrian I, Imig J, Idrissova L, Nastainczyk W, Zeuzem S, Alzari PM, Hartmann RW, Piiper A, Biondi RM: **Allosteric activation of the protein kinase PDK1 with low molecular weight compounds.** *EMBO J* 2006, **25**(23):5469–5480. [http://dx.doi.org/10.1038/sj.emboj.7601416].
52. Hindie V, Stroba A, Zhang H, Lopez-Garcia LA, Idrissova L, Zeuzem S, Hirschberg D, Schaeffer F, Jørgensen TJD, Engel M, Alzari PM, Biondi RM: **Structure and allosteric effects of low-molecular-weight activators on the protein kinase PDK1.** *Nat Chem Biol* 2009, **5**(10):758–764. [http://dx.doi.org/10.1038/nchembio.208].
53. Su DS, Lim JJ, Tinney E, Wan BL, Young MB, Anderson KD, Rudd D, Munshi V, Bahnck C, Felock PJ, Lu M, Lai MT, Touch S, Moyer G, Distefano DJ, Flynn JA, Liang Y, Sanchez R, Prasad S, Yan Y, Perlow-Poehnelt R, Torrent M, Miller M, Vacca JP, Williams TM, Anthony NJ: **Substituted tetrahydroquinolines as potent allosteric inhibitors of reverse transcriptase and its key mutants.** *Bioorg Med Chem Lett* 2009, **19**(17):5119–5123. [http://dx.doi.org/10.1016/j.bmcl.2009.07.031].
54. Mathieu V, Fastrez J, Soumillion P: **Engineering allosteric regulation into the hinge region of a circularly permuted TEM-1 beta-lactamase.** *Protein Eng Des Sel* 2010, **23**(9):699–709. [http://dx.doi.org/10.1093/protein/gzq041].
55. Garvie EI: **Bacterial lactate dehydrogenases.** *Microbiol Rev* 1980, **44:**106–139.
56. Iwata S, Kamata K, Yoshida S, Minowa T, Ohta T: **T and R states in the crystals of bacterial L-lactate dehydrogenase reveal the mechanism for allosteric control.** *Nat Struct Biol* 1994, **1**(3):176–185.
57. Maiti R, Domselaar GHV, Zhang H, Wishart DS: **SuperPose: a simple server for sophisticated structural superposition.** *Nucleic Acids Res*

2004, **32**(Web Server issue):W590–W594. [http://dx.doi.org/10.1093/nar/gkh477].
58. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM: **AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.** *J Biomol NMR* 1996, **8**(4):477–486.
59. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.
60. Tama F, Gadea FX, Marques O, Sanejouand YH: **Building-block approach for determining low-frequency normal modes of macromolecules.** *Proteins* 2000, **41:**1–7.
61. Tirion: **Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis.** *Phys Rev Lett* 1996, **77**(9):1905–1908.
62. Tama F, Sanejouand YH: **Conformational change of proteins arising from normal mode calculations.** *Protein Eng* 2001, **14:**1–6.

doi:10.1186/1471-2105-13-273
Cite this article as: Panjkovich and Daura: Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics* 2012 **13**:273.

Chapter 4

Concluding remarks and perspectives

Allosteric sites, and the allosteric phenomenon as a whole, represent a promising territory for the development of novel therapeutics. Computational and theoretical studies have helped to further understand these systems. However, previous studies mentioned above did not benefit from the larger data set now available at ASD (Huang *et al.*, 2011) and they did not implement a structural conservation measure as explained in the first article (Panjkovich and Daura, 2010). In our first article we found that allosteric sites could be conserved at the structural level, but not necessarily so at the level of sequence or primary structure. By then we had no large data set of structurally known allosteric sites, but recent work carried out during my visit to the laboratory of Dr. Mallur Srivatsan Madhusudhan in Singapore is confirming those findings, which until now have not been backed up with significant data.

In principle, a feature that provides a biological advantage will be conserved through evolution and this conservation should be measurable in terms of sequence conservation, so one would expect allosteric sites to display such signal. However, there may be an explanation for this curious lack of evolutionary conservation in allosteric sites. In the work of Wendell Lim (Good *et al.*, 2011) scaffold proteins were found out to be exploited by evolution to harness allosteric regulation. In a similar way, we found that in many cases allosteric binding pockets are common across the protein family. However, they do not always fulfill regulatory tasks, as seen in the case of L-lactate dehydrogenase. Although we did not mention this in the published articles, the fact that protein pockets which serve as allosteric sites are conserved at the structural level point to the same characteristic of the evolutionary process mentioned by Lim: allosteric properties arise through pre-existing features.

In our second article we estimated on a large set of allosteric proteins the degree to which allosteric properties may be related to changes in protein flexibility and we exploited these findings by implementing a predictive methodology.

Given the short calculation times, the predictive method has been integrated in genomic scale pipelines (see Annex II) and it could be used on systematic annotation of protein structures, a task which would be particular useful in the context of structural genomics (Laskowski and Thornton, 2008).

One of the greatest lessons learned during this work was about the large degree of variety found among allosteric proteins and their mechanisms. Nature seems to exploit particular features on each protein or complex to build up allosteric regulation, generating an extremely wide scenario of mechanisms and peculiarities, which renders large-scale approaches aimed at finding common patterns very difficult. Future work will focus on the study of particular protein families, such as GPCRs, given their pharmacological relevance.

As explained in the introduction, we centered our studies on protein structures and did not include details about ligand molecules and their particular properties. This means that the method could be improved by further exploring such ligand-related features. Furthermore, the method could also incorporate detailed molecular dynamics simulations even at the expense of considerably larger calculation times.

In summary, the objectives layed out at the beginning of this thesis have been accomplished, we hope that the results that have been published, as well as the prediction method which is now available through a web-server, will help researchers in gaining a deeper understanding of allosteric systems and exploiting this knowledge to develop novel therapeutics.

Bibliography

Ackers,G.K. and Holt,J.M. (2006) Asymmetric cooperativity in a symmetric tetramer: human hemoglobin. *J Biol Chem,* **281** (17), 11441–11443.

Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J. (2002) *Molecular Biology of the Cell*. 4th edition,, Garland.

Ascenzi,P., Bocedi,A., Bolli,A., Fasano,M., Notari,S. and Polticelli,F. (2005) Allosteric modulation of monomeric proteins. *Biochem Mol Biol Educ,* **33** (3), 169–176.

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** (1), 235–242.

Bowman,G.R. and Geissler,P.L. (2012) Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc Natl Acad Sci U S A,* **109** (29), 11681–11686.

Cardenas,M.L., Rabajille,E. and Niemeyer,H. (1978) Maintenance of the monomeric structure of glucokinase under reacting conditions. *Arch Biochem Biophys,* **190** (1), 142–148.

Changeux,J.P. (2011) 50th anniversary of the word 'allosteric'. *Protein Sci,* **20** (7), 1119–1124.

Changeux,J.P. (2012) Allostery and the monod-wyman-changeux model after 50 years. *Annu Rev Biophys,* **41**, 103–133.

Changeux,J.P. and Edelstein,S.J. (2005) Allosteric mechanisms of signal transduction. *Science,* **308** (5727), 1424–1428.

Christian Bohr,K.H. and Krogh,A. (1904) Uber einen in biologischer beziehung wichtigen einfluss, den die kohlensaurespannung des blutes auf dessen sauerstoffbindung ubt. *Skand. Arch. Physiol.,* **16**, 401–412.

Conn,P.J., Christopoulos,A. and Lindsley,C.W. (2009) Allosteric modulators of gpcrs: a novel approach for the treatment of cns disorders. *Nat Rev Drug Discov,* **8** (1), 41–54.

Cooper,A. and Dryden,D.T. (1984) Allostery without conformational change. a plausible model. *Eur Biophys J,* **11** (2), 103–109.

Cori GT,C.S. and CF.,C. (1938) The action of nucleotides in the disruptive phosphorylation of glycogen. *Journal of Biological Chemistry,* **123**, 381–389.

Cui,Q. and Karplus,M. (2008) Allostery and cooperativity revisited. *Protein Sci,* **17** (8), 1295–1307.

Daily,M.D. and Gray,J.J. (2007) Local motions in a benchmark of allosteric proteins. *Proteins,* **67** (2), 385–399.

Daily,M.D. and Gray,J.J. (2009) Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput Biol,* **5** (2), e1000293.

DeDecker,B.S. (2000) Allosteric drugs: thinking outside the active-site box. *Chem Biol,* **7** (5), R103–R107.

del Sol,A., Tsai,C.J., Ma,B. and Nussinov,R. (2009) The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure,* **17** (8), 1042–1050.

Demerdash,O.N.A., Daily,M.D. and Mitchell,J.C. (2009) Structure-based predictive models for allosteric hot spots. *PLoS Comput Biol,* **5** (10), e1000531.

Ferreon,A.C.M., Ferreon,J.C., Wright,P.E. and Deniz,A.A. (2013) Modulation of allostery by protein intrinsic disorder. *Nature,* **498** (7454), 390–394.

Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L.L. and Bateman,A. (2008) The pfam protein families database. *Nucleic Acids Res,* **36** (Database issue), D281–D288.

Flynn,T.C., Swint-Kruse,L., Kong,Y., Booth,C., Matthews,K.S. and Ma,J. (2003) Allosteric transition pathways in the lactose repressor protein core domains: asymmetric motions in a homodimer. *Protein Sci,* **12** (11), 2523–2541.

Frauenfelder,H., Sligar,S.G. and Wolynes,P.G. (1991) The energy landscapes and motions of proteins. *Science,* **254** (5038), 1598–1603.

Freire,E. (2000) Can allosteric regulation be predicted from structure? *Proc Natl Acad Sci U S A,* **97** (22), 11680–11682.

Goncearenco,A., Mitternacht,S., Yong,T., Eisenhaber,B., Eisenhaber,F. and Berezovsky,I.N. (2013) Spacer: server for predicting allosteric communication and effects of regulation. *Nucleic Acids Res,* **41** (Web Server issue), W266–W272.

Good,M.C., Zalatan,J.G. and Lim,W.A. (2011) Scaffold proteins: hubs for controlling the flow of cellular information. *Science,* **332** (6030), 680–686.

Goodey,N.M. and Benkovic,S.J. (2008) Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol,* **4** (8), 474–482.

Gunasekaran,K., Ma,B. and Nussinov,R. (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins,* **57** (3), 433–443.

Hardy,J.A., Lam,J., Nguyen,J.T., O'Brien,T. and Wells,J.A. (2004) Discovery of an allosteric site in the caspases. *Proc Natl Acad Sci U S A,* **101** (34), 12461–12466.

Hardy,J.A. and Wells,J.A. (2004) Searching for new allosteric sites in enzymes. *Curr Opin Struct Biol,* **14** (6), 706–715.

Hawkins,R.J. and McLeish,T.C.B. (2004) Coarse-grained model of entropic allostery. *Phys. Rev. Lett.,* **93**, 098104.

Helmstaedt,K., Krappmann,S. and Braus,G.H. (2001) Allosteric regulation of catalytic activity: escherichia coli aspartate transcarbamoylase versus yeast chorismate mutase. *Microbiol Mol Biol Rev,* **65** (3), 404–21, table of contents.

Heredia,V.V., Thomson,J., Nettleton,D. and Sun,S. (2006) Glucose-induced conformational changes in glucokinase mediate allosteric regulation: transient kinetic analysis. *Biochemistry,* **45** (24), 7553–7562.

Hill,A.V. (1910) The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *Journal of Physiology,* **40**, iv–vii.

Hilser,V.J. (2010) Biochemistry. an ensemble view of allostery. *Science,* **327** (5966), 653–654.

Hilser,V.J. (2013) Structural biology: signalling from disordered proteins. *Nature,* **498** (7454), 308–310.

Huang,B. and Schroeder,M. (2006) Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct Biol,* **6**, 19.

Huang,W., Lu,S., Huang,Z., Liu,X., Mou,L., Luo,Y., Zhao,Y., Liu,Y., Chen,Z., Hou,T. and Zhang,J. (2013) Allosite: a method for predicting allosteric sites. *Bioinformatics,* **29** (18), 2357–2359.

Huang,Z., Zhu,L., Cao,Y., Wu,G., Liu,X., Chen,Y., Wang,Q., Shi,T., Zhao,Y., Wang,Y., Li,W., Li,Y., Chen,H., Chen,G. and Zhang,J. (2011) Asd: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Res,* **39** (Database issue), D663–D669.

Kamata,K., Mitsuya,M., Nishimura,T., Eiki,J.I. and Nagata,Y. (2004) Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure,* **12** (3), 429–438.

Kaya,C., Armutlulu,A., Ekesan,S. and Haliloglu,T. (2013) Mcpath: monte carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic Acids Res,* **41** (Web Server issue), W249–W255.

Kern,D. and Zuiderweg,E.R.P. (2003) The role of dynamics in allosteric regulation. *Curr Opin Struct Biol,* **13** (6), 748–757.

Kidd,B.A., Baker,D. and Thomas,W.E. (2009) Computation of conformational coupling in allosteric proteins. *PLoS Comput Biol,* **5** (8), e1000484.

Kohl,A., Amstutz,P., Parizek,P., Binz,H.K., Briand,C., Capitani,G., Forrer,P., Pluckthun,A. and Grutter,M.G. (2005) Allosteric inhibition of aminoglycoside phosphotransferase by a designed ankyrin repeat protein. *Structure,* **13** (8), 1131–1141.

Koshland,D.E., Nemethy,G. and Filmer,D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry,* **5** (1), 365–385.

Kresge,N., Simoni,R.D. and Hill,R.L. (2005) The enzymology of feedback inhibition by arthur b. pardee. *Journal of Biological Chemistry,* **280** (41), e38.

Laskowski,R.A., Gerick,F. and Thornton,J.M. (2009) The structural basis of allosteric regulation in proteins. *FEBS Lett,* **583** (11), 1692–1698.

Laskowski,R.A. and Thornton,J.M. (2008) Understanding the molecular machinery of genetics through 3D structures. *Nat Rev Genet,* **9** (2), 141–151.

Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol,* **257** (2), 342–358.

Liu,J. and Nussinov,R. (2008) Allosteric effects in the marginally stable von hippel-lindau tumor suppressor protein and allostery-based rescue mutant design. *Proc Natl Acad Sci U S A,* **105** (3), 901–906.

Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science,* **286** (5438), 295–299.

Maita,N., Okada,K., Hatakeyama,K. and Hakoshima,T. (2002) Crystal structure of the stimulatory complex of gtp cyclohydrolase i and its feedback regulatory protein gfrp. *Proc Natl Acad Sci U S A,* **99** (3), 1212–1217.

McCammon,J.A., Gelin,B.R. and Karplus,M. (1977) Dynamics of folded proteins. *Nature,* **267** (5612), 585–590.

Ming,D. and Wall,M.E. (2005) Quantifying allosteric effects in proteins. *Proteins,* **59** (4), 697–707.

Ming,D. and Wall,M.E. (2006) Interactions in native binding sites cause a large change in protein dynamics. *J Mol Biol,* **358** (1), 213–223.

Mistry,J., Bateman,A. and Finn,R.D. (2007) Predicting active site residue annotations in the pfam database. *BMC Bioinformatics,* **8**, 298.

Mitternacht,S. and Berezovsky,I.N. (2011) Binding leverage as a molecular basis for allosteric regulation. *PLoS Comput Biol,* **7** (9), e1002148.

Moehler,H., Fritschy,J.M. and Rudolph,U. (2002) A new benzodiazepine pharmacology. *J Pharmacol Exp Ther,* **300** (1), 2–8.

Monod,J., Changeux,J.P. and Jacob,F. (1963) Allosteric proteins and cellular control systems. *J Mol Biol,* **6**, 306–329.

Monod,J. and Jacob,F. (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb Symp Quant Biol,* **26**, 389–401.

Monod,J., Wyman,J. and Changeux,J.P. (1965) On the nature of allosteric transitions: a plausible model. *J Mol Biol,* **12**, 88–118.

Nussinov,R. and Tsai,C.J. (2013) Allostery in disease and in drug discovery. *Cell,* **153** (2), 293–305.

Nussinov,R., Tsai,C.J. and Csermely,P. (2011) Allo-network drugs: harnessing allostery in cellular networks. *Trends Pharmacol Sci,* **0**, 0.

Ortiz,A.R., Strauss,C.E.M. and Olmea,O. (2002) Mammoth (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci,* **11** (11), 2606–2621.

Ota,N. and Agard,D.A. (2005) Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J Mol Biol,* **351** (2), 345–354.

Panjkovich,A. and Daura,X. (2010) Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC Struct Biol,* **10**, 9.

Panjkovich,A. and Daura,X. (2012) Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics,* **13** (1), 273.

Pauling,L. (1935) The oxygen equilibrium of hemoglobin and its structural interpretation. *Proc Natl Acad Sci U S A,* **21** (4), 186–191.

Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein-protein interaction. *J Mol Biol,* **271** (4), 511–523.

Peracchi,A. and Mozzarelli,A. (2011) Exploring and exploiting allostery: models, evolution, and drug targeting. *Biochim Biophys Acta,* **1814** (8), 922–933.

Perutz,M.F. (1970) Stereochemistry of cooperative effects in haemoglobin. *Nature,* **228** (5273), 726–739.

Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res,* **32** (Database issue), D129–D133.

Raddatz,R., Schaffhauser,H. and Marino,M.J. (2007) Allosteric approaches to the targeting of g-protein-coupled receptors for novel drug discovery: a critical assessment. *Biochem Pharmacol,* **74** (3), 383–391.

Rader,A.J. and Brown,S.M. (2011) Correlating allostery with rigidity. *Mol Biosyst,* **7** (2), 464–471.

Reichard,P. (2010) Ribonucleotide reductases: substrate specificity by allostery. *Biochem Biophys Res Commun,* **396** (1), 19–23.

Rodriguez,G.J., Yao,R., Lichtarge,O. and Wensel,T.G. (2010) Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc Natl Acad Sci U S A,* **107** (17), 7787–7792.

Schneider,T.D. (2000) Evolution of biological information. *Nucleic Acids Res,* **28** (14), 2794–2799.

Stadtman,E.R., Cohen,G.N., Bras,G.L. and de Robichon-Szulmajster (1961) Selective feedback inhibition and repression of two aspartokinases in the metabolism of escherichia coli. *Cold Spring Harb Symp Quant Biol,* **26**, 319–321.

Tsai,C.J., del Sol,A. and Nussinov,R. (2008) Allostery: absence of a change in shape does not imply that allostery is not at play. *J Mol Biol,* **378** (1), 1–11.

Tsai,C.J., Sol,A.D. and Nussinov,R. (2009) Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Mol Biosyst,* **5** (3), 207–216.

Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature,* **171** (4356), 737–738.

Weber,G. (1972) Ligand binding and internal equilibria in proteins. *Biochemistry,* **11** (5), 864–878.

Yates,R.A. and Pardee,A.B. (1956) Control of pyrimidine biosynthesis in escherichia coli by a feed-back mechanism. *J Biol Chem,* **221** (2), 757–770.

Zorn,J.A. and Wells,J.A. (2010) Turning enzymes on with small molecules. *Nat Chem Biol,* **6** (3), 179–188.

Chapter 5

Supplementary Material

5.1 Annex I - PARS: a web-server for the prediction of Protein Allosteric and Regulatory Sites

5.2 Annex II - antibacTR: dynamic antibacterial-drug-target ranking integrating comparative genomics, structural analysis and experimental annotation

PARS: a web-server for the prediction of Protein Allosteric and Regulatory Sites

Alejandro Panjkovich, Xavier Daura

Abstract

Protein activity regulation is a key aspect of life. Unveiling its details at the atomic level is key to understanding signalling and metabolic pathways among other fundamental processes. The most common and powerful way of regulating protein function is allostery. It has been increasingly calling the attention of medicinal chemists given the potential it possesses for the discovery of novel therapeutics. In this context, PARS is a simple and quick method that queries protein dynamics and structural conservation to readily identify pockets on a protein structure which may exert a regulatory effect upon the binding of a small-molecule ligand. PARS is freely available as a web-server at http://bioinf.uab.cat/pars

Introduction

Tight regulation of protein function is fundamental to life. Proteins involved in metabolic pathways, signalling cascades and genomic transcription among other processes within the living cell are commonly under allosteric regulation, *i.e.* their activity is modified through the binding of a ligand molecule to the protein in a site different from the active site. In fact, allostery is one of the most powerful protein-function regulation mechanisms as it allows proteins to sense and immediately respond to changes in their environment (Monod and Jacob, 1961; Fenton, 2008). Accordingly, the traditional drug-design approach focused on active or primary binding sites can be extended by exploiting allosteric sites, as shown by current efforts on GPCRs (Melancon *et al.*, 2012) or farnesyl pyrophosphate synthase (Jahnke *et al.*, 2010). An advantage of targeting allosteric sites therapeutically is a reduced risk of secondary adverse effects. This is because allosteric sites appear to be significantly less conserved than active sites across homolog proteins (Waelbroeck, 2003), enabling the design of allosteric drugs with high specificity for a single protein within a family. This observation has motivated the development of allosteric drugs for the regulation of phosphodiesterase 4D (PDE4D), for which active-site inhibitors cause emesis, a dose-limiting side effect (Burgin *et al.*, 2010). Moreover, a drug-discovery approach based on allosteric sites may result in the development of not only novel drug-like inhibitors, but activators as well (Peracchi and Mozzarelli, 2011). Other concepts such as the notion of serendipitous allosteric sites, which have no known ligand in nature but can become functional in the presence of an "opportunistic" ligand (Hardy and Wells, 2004) or the idea that any dynamic protein has the potential to be regulated allosterically (Gunasekaran *et al.*, 2004) also contribute to the current level of attention to allosteric sites. In this context, it is to be expected that a deeper understanding of the properties of allosteric sites and their identification would help streamlining the design and discovery of novel therapeutic drugs (Nussinov and Tsai, 2013).

Despite growing efforts, the atomic-level details that explain the functional relationship between distant sites in the same protein molecule have not been elucidated for most of the known cases of allostery (Peracchi and Mozzarelli, 2011). This has motivated several studies aimed

at modelling or computationally predicting the relationship between allosteric and active sites (Lockless and Ranganathan, 1999; Ming and Wall, 2005; Balabin *et al.*, 2009; Demerdash *et al.*, 2009; Kidd *et al.*, 2009; Daily and Gray, 2009; Mitternacht and Berezovsky, 2011; England, 2011; Reynolds *et al.*, 2011; Weinkam *et al.*, 2012), each of these studies representing a substantial step forward in the understanding of allostery. Studying allostery at a larger scale has been however difficult until the publication in 2011 of the AlloSteric Database (ASD), the first public initiative to organize knowledge on allosteric sites (Huang *et al.*, 2011). We have used ASD data to benchmark a novel method that integrates protein structural conservation and flexibility to predict the location of allosteric and/or regulatory sites, obtaining a positive predictive value of 65% when using strict parameters (Panjkovich and Daura, 2012). In this applications note we describe an easy-to-use web-server that makes the methodology available to the scientific community at "http://bioinf.uab.cat/pars".

Description

The current perspective on allosteric transitions and associated regulatory events relies on the "population shift" concept (Cui and Karplus, 2008). Briefly, the protein or protein complex explores different conformations (both active and inactive) in solution and the allosteric ligand "shifts" the population or ensemble of conformations upon binding, effectively modulating the protein's activity rate (Kar *et al.*, 2010). The conformational space explored by the protein can be sampled using computational methods such as molecular dynamics (Chiappori *et al.*, 2012; Pandini *et al.*, 2012) or normal mode analysis (NMA) (Dykeman and Twarock, 2010). In our approach, protein dynamics are queried through NMA allowing fast large-scale (potentially genome-wide) analyses. The method developed previously (Panjkovich and Daura, 2012) is now implemented as a web-server working as follows: (1) Initially the user uploads a protein-structure file (PDB format) and selects which chains and ligands should be considered for the calculations. (2) Once the job is submitted, the protein surface is analyzed to predict putative ligand-binding sites, where a simplified representation of a small-molecule is placed to simulate the presence of a ligand. The prediction of putative ligand-binding sites can be turned off if the user is interested in scanning only sites which are already occupied by a ligand in the protein structure. (3) NMA is carried out for the apo structure (without ligands). (4) For each of the ligands, both native or simplified representation, and each of the potential binding sites, a NMA is executed for the protein-ligand complex. (5) If a significant difference is found between the normal modes of the apo and ligand-bound states of the protein, the binding site is marked as potentially allosteric.

Complementarily, if enough structural data is available for the protein family, the structural conservation of each pocket is also measured. We have previously shown that allosteric sites tend to be structurally conserved within a protein family (Panjkovich and Daura, 2010) and that the incorporation of this measure improves the capacity of the method based on dynamics to correctly identify allosteric sites (Panjkovich and Daura, 2012). Immediately after submission the user is provided with a link where results can be accessed once the calculation has finished. Optionally, the user can choose to be notified by e-mail when results are available. A standard run on a 300 residues protein will take around 2 minutes. Even though calculation

times increase exponentially with the number of residues, larger systems can be processed within reasonable time (*e.g.* 500 residues $\approx$ 6 minutes and 1000 residues $\approx$ 45 minutes). The results are delivered as a table and, if the browser allows it, the binding-sites can be explored on-line on the three-dimensional protein structure by means of the Jmol package (http://www.jmol.org/). Binding sites and ligands are color-coded according to given thresholds for the level of structural conservation ($> 50\%$) and the predicted effect on protein dynamics ($p\text{-}value < 0.05$). Both a protein structure file including the position of identified sites and a table can be downloaded as well for further processing.

We expect that this easy-to-use and relatively fast web-server will prove useful for medicinal chemists and other researchers studying the regulation of protein function for biochemical characterization and other applied tasks such as binding-site prioritization for virtual drug screening.

Acknowledgements

This project is supported by funding under the Seventh Research Framework Programme of the European Union (ref. HEALTH-F3-2009-223101). AP acknowledges the FPU Scholarship from MICINN, Spanish Gov.

Bibliography

Balabin,I.A., Yang,W. and Beratan,D.N. (2009) Coarse-grained modeling of allosteric regulation in protein receptors. *Proc Natl Acad Sci U S A,* **106** (34), 14253–14258.

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. and Gurney,M.E. (2010) Design of phosphodiesterase 4d (pde4d) allosteric modulators for enhancing cognition with improved safety. *Nat Biotechnol,* **28** (1), 63–70.

Chiappori,F., Merelli,I., Colombo,G., Milanesi,L. and Morra,G. (2012) Molecular mechanism of allosteric communication in hsp70 revealed by molecular dynamics simulations. *PLoS Comput Biol,* **8** (12), e1002844.

Cui,Q. and Karplus,M. (2008) Allostery and cooperativity revisited. *Protein Sci,* **17** (8), 1295–1307.

Daily,M.D. and Gray,J.J. (2009) Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput Biol,* **5** (2), e1000293.

Demerdash,O.N.A., Daily,M.D. and Mitchell,J.C. (2009) Structure-based predictive models for allosteric hot spots. *PLoS Comput Biol,* **5** (10), e1000531.

Dykeman,E.C. and Twarock,R. (2010) All-atom normal-mode analysis reveals an rna-induced allostery in a bacteriophage coat protein. *Phys Rev E Stat Nonlin Soft Matter Phys,* **81** (3 Pt 1), 031908.

England,J.L. (2011) Allostery in protein domains reflects a balance of steric and hydrophobic effects. *Structure,* **19** (7), 967–975.

Fenton,A.W. (2008) Allostery: an illustrated definition for the 'second secret of life'. *Trends Biochem Sci,* **33** (9), 420–425.

Gunasekaran,K., Ma,B. and Nussinov,R. (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins,* **57** (3), 433–443.

Hardy,J.A. and Wells,J.A. (2004) Searching for new allosteric sites in enzymes. *Curr Opin Struct Biol,* **14** (6), 706–715.

Huang,Z., Zhu,L., Cao,Y., Wu,G., Liu,X., Chen,Y., Wang,Q., Shi,T., Zhao,Y., Wang,Y., Li,W., Li,Y., Chen,H., Chen,G. and Zhang,J. (2011) Asd: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Res,* **39** (Database issue), D663–D669.

Jahnke,W., Rondeau,J.M., Cotesta,S., Marzinzik,A., Pell,X., Geiser,M., Strauss,A., Gtte,M., Bitsch,F., Hemmig,R., Henry,C., Lehmann,S., Glickman,J.F., Roddy,T.P., Stout,S.J. and Green,J.R. (2010) Allosteric non-bisphosphonate fpps inhibitors identified by fragment-based discovery. *Nat Chem Biol,* **6** (9), 660–666.

Kar,G., Keskin,O., Gursoy,A. and Nussinov,R. (2010) Allostery and population shift in drug discovery. *Curr Opin Pharmacol,* **10** (6), 715–722.

Kidd,B.A., Baker,D. and Thomas,W.E. (2009) Computation of conformational coupling in allosteric proteins. *PLoS Comput Biol,* **5** (8), e1000484.

Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science,* **286** (5438), 295–299.

Melancon,B.J., Hopkins,C.R., Wood,M.R., Emmitte,K.A., Niswender,C.M., Christopoulos,A., Conn,P.J. and Lindsley,C.W. (2012) Allosteric modulation of seven transmembrane spanning receptors: theory, practice, and opportunities for central nervous system drug discovery. *J Med Chem,* **55** (4), 1445–1464.

Ming,D. and Wall,M.E. (2005) Quantifying allosteric effects in proteins. *Proteins,* **59** (4), 697–707.

Mitternacht,S. and Berezovsky,I.N. (2011) Binding leverage as a molecular basis for allosteric regulation. *PLoS Comput Biol,* **7** (9), e1002148.

Monod,J. and Jacob,F. (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb Symp Quant Biol,* **26**, 389–401.

Nussinov,R. and Tsai,C.J. (2013) Allostery in disease and in drug discovery. *Cell,* **153** (2), 293–305.

Pandini,A., Fornili,A., Fraternali,F. and Kleinjung,J. (2012) Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J,* **26** (2), 868–881.

Panjkovich,A. and Daura,X. (2010) Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC Struct Biol,* **10**, 9.

Panjkovich,A. and Daura,X. (2012) Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics,* **13** (1), 273.

Peracchi,A. and Mozzarelli,A. (2011) Exploring and exploiting allostery: models, evolution, and drug targeting. *Biochim Biophys Acta,* **1814** (8), 922–933.

Reynolds,K.A., McLaughlin,R.N. and Ranganathan,R. (2011) Hot spots for allosteric regulation on protein surfaces. *Cell,* **147** (7), 1564–1575.

Waelbroeck,M. (2003) Allosteric drugs acting at muscarinic acetylcholine receptors. *Neurochem Res,* **28** (3-4), 419–422.

Weinkam,P., Pons,J. and Sali,A. (2012) Structure-based model of allostery predicts coupling between distant sites. *Proc Natl Acad Sci U S A,* **109** (13), 4875–4880.

antibacTR: dynamic antibacterial-drug-target ranking integrating comparative genomics, structural analysis and experimental annotation

Alejandro Panjkovich, Isidre Gibert, Xavier Daura

Abstract

Development of novel antibacterial drugs is both an urgent healthcare necessity and a partially neglected field. The last decades have seen a substantial decrease in the discovery of novel antibiotics, which combined with the recent thrive of multi-drug-resistant pathogens have generated a scenario of general concern. The procedures involved in the discovery and development of novel antibiotics are economically challenging, time consuming and lack any warranty of success. Furthermore, the return-on-investment for an antibacterial drug is usually marginal when compared to other therapeutics, which in part explains the decrease of private investment. In this work we present antibacTR, a computational pipeline designed to aid researchers in the selection of potential drug targets, one of the initial steps in antibacterial-drug discovery. The approach was designed and implemented as part of two publicly funded initiatives aimed at discovering novel antibacterial targets, mechanisms and drugs for a priority list of Gram-negative pathogens: *Acinetobacter baumannii*, *Escherichia coli*, *Helicobacter pylori*, *Pseudomonas aeruginosa* and *Stenotrophomonas maltophilia*. Future releases will extend this list to additional multi-drug resistant Gram-negative pathogens of clinical relevance. antibacTR is based on sequence comparisons and queries to multiple databases (*e.g.* gene essentiality, virulence factors) to rank proteins according to their potential as antibacterial targets. The dynamic ranking of potential drug targets can easily be executed, customized and accessed by the user through a web interface which also integrates computational analyses performed in-house and visualizable on-site. These include three-dimensional modeling of protein structures and prediction of active sites among other functionally relevant ligand-binding sites. Versatility and ease-of-use have been emphasized so that this tool may effectively assist microbiologists, medicinal-chemists and other researchers working in the field of antibacterial drug-discovery. The public web-interface for antibacTR is available at 'http://bioinf.uab.cat/antibactr'.

Introduction

Since their initial discovery and application during the early 20th century, antibiotics have been playing a key role in public health worldwide. These 'miracle drugs' have contributed significantly to the increase in life expectancy since the end of World War II. Besides curing infections, they also prevent amputations and blindness and are involved in multiple healthcare procedures such as joint-replacement, surgery, new cancer treatments, etc (Arias and Murray, 2009). However, after peaking during the 1960's, the discovery of new antibiotics has fallen off dramatically. The present scarcity of novel antibiotics becomes a major health concern in light of the remarkable ability of bacteria to rapidly evolve resistance mechanisms which erode the therapeutic effect of known antibiotics (Baquero *et al.*, 2009). Nowadays multi-drug-resistant bacterial infections are increasing in both developing and developed countries and in both com-

munity and nosocomial settings (Boucher *et al.*, 2009). It has been reported that a number of pathogens, including *Staphylococcus aureus*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, *Acinetobacter baumannii* and some Enterobacteriaceae have developed resistance to a wide range of antimicrobial agents at an alarming rise, with some strains becoming truly pan-resistant (Council, 2007; Souli *et al.*, 2008). However, pharmaceutical companies have not been investing on the development of new antibacterial drugs with corresponding efforts, mainly due to economic criteria that favour other therapeutic areas with better return-on-investment ratios (Talbot *et al.*, 2006; Payne *et al.*, 2007). The few antibacterial agents that have been launched during the last decade (*e.g.* linezolid, daptomycin) have a good activity against Gram-positive bacteria such as methicillin-resistant *S. aureus* (MRSA) and vancomycin-resistant enterococci (Tally and DeBruin, 2000). However, cases of resistance for these new Gram-positive antibiotics have been reported recently as well (van Hal and Paterson, 2011).

The situation is worse for Gram-negative bacteria, such as *P. aeruginosa* and *A. baumannii*, which are common among nosocomial infections (Bereket *et al.*, 2012) and for which no new antibiotics have reached advanced stages of development (Arias and Murray, 2009). In addition, with the increase in the prevalence of extended spectrum β-lactamase (ESBL)-producing Enterobacteriaceae, the use of carbapenems, a potential alternative to treat infections caused by these microorganisms, is leading to the emergence of multi-drug-resistant Enterobacteriaceae including resistance to carbapenems (Akova *et al.*, 2012).

This scenario emphasizes the relevance of initiatives focused on the discovery of novel targets and antibacterials for combating Gram-negative pathogens. Here, we describe a tool (antibacTR: antibacterial Target Ranking) to support the initial stages of selection of potential antibacterial-drug targets, developed within the context of two such initiatives. antibacTR integrates a database with a pipeline that ranks and filters proteins according to a set of criteria commonly associated to antibacterial targets. The approach is based on protein sequence comparisons, for which we developed an unbiased measure described in the methods section.

The interface used to interrogate the database and access the results has the form of a web-based tool, which has been developed following the suggestions of the experimentalists involved in the two target-discovery initiatives. It includes access to thousands of three-dimensional protein-structure models that can be visualized and downloaded for further analysis through the web-interface. To further exploit the structural models, we incorporated a predictive approach that evaluates putative ligand-binding pockets in terms of their potential to affect protein function upon ligand binding (Panjkovich and Daura, 2010; Panjkovich and Daura, 2012). Links to DrugBank (Wishart *et al.*, 2008) and the Virulence Factors DataBase (VFDB) (Yang *et al.*, 2008), as well as predictions of active-site residues are provided as well. It should be noted that the final aim of the tool is both to rank proteins according to the chosen set of criteria (with weights defined by the user) and to provide for each protein in the ranked list information that could be relevant to antibacterial-drug-target selection. Clearly, the drug-target property is the result of a complex combination of a variable number of non-universal factors, some of which having opposite sign for different types of targets or mechanisms. Thus, this is a tool to support target discovery efforts, not a target-prediction tool. In other words, it will not safe the user from scanning and evaluating a large number of proteins, it will simply provide him/her with additional means to do it.

All measurements and predictions are pre-computed, which allows the application to return

full rankings and links to the relevant information within seconds. This characteristic distinguishes our tool from similar ones such as the UniDrug-Target (UDT) database, which can be used to perform comparative analyses online with computation at time of request (Chanumolu *et al.*, 2012). Besides execution speed, our approach differs from UDT and related ones such as the Prokaryotic-genome Analysis Tool (PGAT) (Brittnacher *et al.*, 2011) in its focus. While these tools succeed at providing comparative-analysis means that can be used through a web-interface, our dynamic approach focuses on speed, ease of use and an integrative solution that allows the user to quickly scan putative antibacterial targets and relevant information such as three-dimensional structural models and other predictions, while the comparative analysis is just one of the underlying features.

This methodology was originally developed for two specific projects and their target bacteria, including three of the big four Gram-negative pathogens in relation to systemic infections, i.e. *Escherichia coli*, *Klebsiella pneumoniae*, *A. baumannii* and *P. aeruginosa* (Council, 2007). However, we have now expanded the coverage to include all sequenced Gram-negative pathogens (*i.e.* at present, a total of 74 species), this should make the system useful to many more researchers in the field. Through this article, we describe the approach and make the web-based tool publicly available.

Results

Computational target-ranking pipeline

Typically, one of the first steps in a target-discovery project is to readily select, among thousands of proteins composing the pathogens' proteomes, those with the highest chance of becoming useful therapeutic targets. Following the lines defined by previous studies (White and Kell, 2004), we developed an algorithm to score and rank potential drug targets in pathogenic organisms by evaluating a modular set of criteria that are commonplace in antimicrobial-development efforts (Payne *et al.*, 2007): 1) the presence of the protein in different pathogens, 2) evolutionary conservation, 3) essentiality, 4) presence of isoforms and paralogs in the proteome, 5) similarity to human proteins. We implemented a set of five weighted scores that cover these criteria and defined a scoring function combining them.

The first two concepts were incorporated as two independent scores, measuring the conservation of the protein among Gram-negative organisms and among different strains of the same species, respectively. Conservation among strains is a basic requirement for target consideration. Conservation among Gram-negative species is highly desirable as it enables the development of broad-spectrum solutions and increases economic viability. In addition, well conserved targets will presumably have low tolerance to mutations, decreasing the chance of resistance to emerge by this type of mechanism.

Essential proteins, which inhibition compromises bacterial viability, are potential antibacterial targets by definition. We implemented a binary score by marking genes known to be essential from previous experimental work (Zhang and Zhang, 2008).

The remaining two scores are given negative weights. If the protein under consideration has isoforms and/or paralogs the pathogen may readily develop resistance by functional substitution, and the effect of the antibacterial may be also reduced by competitive binding to non-essential

forms. We considered similarity to human proteins negative as well, since close human homologs to the target may interact with the drug, giving rise to unwanted side-effects.

The scoring and ranking scheme, partially following the work of White and Kell (White and Kell, 2004) provides an advantage when compared to static selection or filtering approaches (Sakharkar *et al.*, 2004; Chanumolu *et al.*, 2012). In our case, if further experimental analysis reveals that a given protein is not suitable as a drug target, work can continue with the next protein in the ranking. Moreover, it would be straightforward to incorporate new criteria into the ranking scheme if needed.

The pipeline to which each proteome of interest was subjected is illustrated in Figure 1, which summarizes the approach.

Sequence-based analysis

Currently, the database covers 74 Gram-negative pathogens, including 224 distinct strains. For pathogens distinguished by their prevalence in community and/or nosocomial infections and the incidence of drug-resistant isolates (Boucher *et al.*, 2009; Bereket *et al.*, 2012): *Acinetobacter baumannii*, *Escherichia coli*, *Helicobacter pylori*, *Pseudomonas aeruginosa* and *Stenotrophomonas maltophilia* we included all fully sequenced available strains (82 strains, which conform a 'priority set'). For the rest of the species, we included all strains that were marked as 'human pathogens' in HAMAP (142 strains) (Pedruzzi *et al.*, 2013). This query data set was compared against the human proteome and a reference set of 770 Gram-negative proteomes (494 distinct species), by means of the BLAST program (Altschul *et al.*, 1997) using default parameters. BLAST searches are very fast, however resulting E-values depend on the alignment itself and on other parameters such as the size of the database scanned. We needed unbiased similarity scores between proteins matched during the sequence-based search to keep results valid in case of further increasing the size of the data sets. To attain this objective, we further aligned BLAST matches (E-value <= 0.0001) using the Smith-Waterman algorithm and calculated 'normalized sequence similarity scores' (NS). NS values were then used to pre-compute toxicity, presence of isoforms or paralogs and the two conservation scores for each protein in the query data set, as described in further detail in the Methods section.

Queries to external databases

Besides comparative analysis, a round of database queries was also performed to integrate additional information. Thus, protein sequences were compared to the list of known drug-targets available at the DrugBank database (Wishart *et al.*, 2008) and sequence-based searches were also performed against virulence factors available at the Virulence Factors DataBase (VFDB) (Yang *et al.*, 2008). Out of the total 777,585 proteins in the query data set, 375,016 matched a known target in DrugBank (38%) and 193,135 proteins matched a known virulence factor at VFDB (25%). This information was not incorporated as ranking scores, but it is available through the web-interface described below for researchers to evaluate themselves the relevance of such matches in each particular case.

Three-dimensional homology modeling

Researchers evaluating prospective drug targets may benefit from the availability of protein structural data. For the organisms in the priority set, we performed a large-scale homology modeling of all protein sequences for which we found valid structural templates as explained in the Methods section. In total, we generated three-dimensional homology models for 136,141 proteins (covering 47% of the priority set). This number was obtained after discarding models presenting less than 30% sequence identity (target-template) or G-factors below -1.00 (Laskowski *et al.*, 1996). All models were generated by means of the MODELLER program (Eswar *et al.*, 2008) using default parameters.

To save computational power, proteins belonging to other strains were not modeled automatically. However, if the user is interested in obtaining one of such homology models, we have implemented an option at the web-interface for automatic submission of the selected modeling task.

Active-site prediction

To further add relevant information on putative targets, we applied a sequence-based approach (Mistry *et al.*, 2007) to predict the location of active-site residues. The method is based on comparing query sequences to homologs for which the position of the active site has been annotated. After analyzing the whole query set (777,585 proteins), this procedure predicted the location of active-site residues for 90,482 proteins (11.6%). Proteins with a predicted active site display a link to the details of the prediction in the web interface described below.

Pocket analysis

For proteins for which we could build a three-dimensional homology model, we predicted the location of ligand-binding sites on the structure by means of the LIGSITEcs program (Huang and Schroeder, 2006). We further analyzed the ligand-binding sites using two previously developed methodologies which estimate the regulatory potential of particular ligand-binding pockets. When possible, the structural conservation of predicted pockets was measured considering the evolutionary record of the protein family, given that conserved pockets may have a relevant biological role (Panjkovich and Daura, 2010). Furthermore, using Normal Mode Analysis we estimated the effect of ligand binding upon overall protein flexibility, a measure which has been used in combination with structural conservation to predict the location of allosteric sites (Panjkovich and Daura, 2012). As described below, the user can visualize the protein structure and predictions online.

Interface and access to results

Interactive access to results is available through the web-interface at 'http://bioinf.uab.cat/antibactr'.

This interface allows the user to select the organisms and strain of interest, set custom weights to the different scores and then proceed to calculate the corresponding ranking. If the user wishes to ignore a specific ranking parameter, a weight of 0 (zero) can be applied. The system has been built in such a way that normalization of scores is performed only among the

selected set of strains and parameters. To further facilitate the analysis of results, the user may also limit the amount of top-ranked entries that are displayed. Once the ranking procedure is finished (it takes a few seconds), the ranking is printed to the browser. An option is available for downloading the ranking to the local computer in tab-delimited text format, useful for researchers interested in further processing the data. Targets are displayed in ranked order and individual scores are shown for each protein after normalization but prior to weighting. A brief description of the biological function is displayed for each protein but, to facilitate immediate access to full annotation and other relevant data, a link to the related Uniprot entry is provided as well (Consortium, 2009). In cases where the target shows sequence similarity to an already known drug target or virulence factor, the corresponding links are also provided. In addition, specific links with details on predicted active sites and homology models are given. If a homology model is supplied, the user may download model coordinates in PDB format and target-template alignments generated during the modeling process, along with sequence identity, DOPE score and other relevant modeling data (Eswar *et al.*, 2008). Furthermore, available protein structures can be visualized using *Jmol* (http://www.jmol.org) along with the results of the pocket analysis previously described (Panjkovich and Daura, 2012).

User query sequences

Besides the ranking of complete proteomes, researchers may want to look at the ranking of a few selected proteins of their particular interest. To achieve this functionality, we added the possibility to include the user's own query sequences in an optional field. These sequences are then compared by means of the BLAST program against our query data set (224 strains). Scoring and ranking proceed as normally, but results are then displayed only for significant hits within our data set. Details of this BLAST search are also available to the user.

Discussion

Large-scale comparison of organisms at the genome level is a technique common to many fields of biology and medicine. In the past years complex approaches involving phylogenetic and metabolic studies have been published (Fournier *et al.*, 2006; Lee *et al.*, 2009). However, comparative genomics initiatives in drug discovery have been criticized for their limited success in finding new active compounds (Mills, 2006; Coates and Hu, 2007). Yet, comparative genomics and proteomics continue to shed light on the workings of bacterial drug-resistance and virulence (Kos *et al.*, 2012; Piras *et al.*, 2012).

Far from attempting to solve the problem of target identification in one strike, our motivation was to implement a straightforward computational approach that would prove useful as an initial filtering and ranking step, aiding researchers in the quest for novel drug targets.

At the time of this writing no equivalent tool to the one presented here is available, however a few servers provide slightly related functionalities and could be used in a complementary fashion. For example the Prokaryotic-genome Analysis Tool (PGAT), developed by Brittnacher and collaborators is a general comparative genomics tool focused particularly in comparing different strains of the same species (Brittnacher *et al.*, 2011). PGAT allows the user to carry out a series of interesting analyses including information on metabolic pathways, but it does not

provide specific drug-target related information, unlike the UniDrug-Target (UDT) database which clearly focuses on that aspect (Chanumolu *et al.*, 2012). The latter presents candidate targets as proteins which are present in pathogenic bacteria but absent in commensal strains. This is a reasonable approach which in our tool can be achieved by setting a negative value for the strain conservation score and it is also one of the functions available at PGAT. However, given its focus on pathogen-specific proteins UDT's approach tends to discard evolutionary conserved proteins, such as many well known broad-spectrum targets (Imming *et al.*, 2006).

The amount of well known and characterized protein drug targets is currently in the order of hundreds (Imming *et al.*, 2006). To illustrate the potential of the tool presented here, we provide a few examples of already known antibacterial targets. Certain proteins involved in the replication of DNA are targeted by fluoroquinolones such as Ciprofloxacin (Lee *et al.*, 2005; Imming *et al.*, 2006). For example, Ciprofloxacin targets DNA topoisomerase 4 subunit B and DNA gyrase subunit A. Even though resistance to fluoroquinolones has been observed in pathogens with mutations in these proteins, new compounds with antibacterial activity on the resistant strains are being developed by studying these targets (Bax *et al.*, 2010). These proteins appear on the top 3% of the full proteome ranking for *Escherichia coli* K12 (positions 68 and 158, respectively) when we build the ranking using default parameters. This is because both proteins are essential, show high levels of evolutionary conservation but low similarity to human proteins (minimal potential toxicity) and present no isoforms or paralogs according to our pipeline parameters.

Beyond filtering and ranking targets, researchers can also gain insight into potential targets through the structural analysis methods we have implemented into the tool. For example, peptide deformylase [Swiss-Prot:Q9I7A8] from *Pseudomonas aeruginosa* is an essential protein targeted by the antibiotic Actinonin (Guilloteau *et al.*, 2002). This protein is ranked in position 81 among the complete *Pseudomonas aeruginosa* proteome (top 2%) when using default parameters. Our pipeline automatically builds three-dimensional homology models when possible, it then predicts putative ligand-binding sites and evaluates their potential to regulate protein activity as described previously (Panjkovich and Daura, 2012). In this case, the structural analysis (which is pre-calculated and available through the web-interface) predicts one of the putative ligand-binding sites to significantly affect protein flexibility as shown in Figure 2. When we superimpose the automatically generated homology model with the known structure of the protein bound to the antibiotic ([PDB:1LRY] RMSD 0.5) the position of the cavity predicted to be significant matches precisely the location of the antibiotic molecule. This cavity is also considered relevant from an evolutionary perspective, as it shows 100% of structural conservation within its domain family according to the corresponding automatic analysis (Panjkovich and Daura, 2010). The structural conservation of this pocket was to be expected, since it is the protein's active site. Briefly, this case illustrates how in the situation of a poorly characterized protein, our automatic structural analysis may pinpoint not only the potential of the protein as a drug target, but the precise location of the drug-binding pocket.

Another interesting example is 3-oxoacyl-[acyl-carrier-protein] synthase III, which is involved in fatty-acid synthesis. This protein is targeted by Cerulenin, with antifungal effects, and it is a potential antibacterial target as well (Khandekar *et al.*, 2003; Nie *et al.*, 2005; Zhang *et al.*, 2012). It is ranked by our pipeline at position 49 (top 2%) among the full proteome of *Escherichia coli* K12 because it is essential and well conserved, showing in principle no toxicity (similarity to

human proteins). Furthermore, structural analysis reveals one single pocket that could affect the protein's function by perturbing its overall flexibility, as shown in Figure 3. When we evaluate the structural prediction performed on the homology model by superimposing the known structure of the inhibitor-bound protein, we observe that in this case the match of the cavity's geometric center is not as precise as in the previous example of peptide deformylase. Nevertheless, visual inspection shows that the large cavity indicated by the pocket structural analysis is indeed occupied by the inhibitor, so that the automatic procedure would again be pointing the researcher in the right direction, even if no inhibition information would have been available *a priori*.

Of course, multiple other factors beyond the reach of a mere computational approach participate in defining a protein as a good antibacterial target. Indeed the final outcome of clinical trials can hardly be predicted (Mills, 2006). However, we considered relevant to infer how well the pipeline presented in this work ranks already known targets. We gathered all known targets of 'approved' drugs, as annotated in DrugBank (Wishart *et al.*, 2008), that belonged to any of the strains analyzed in this work. We found a total of 57 proteins identified through their Uniprot ID (Consortium, 2009). Out of those 57, a majority (48) belong to *Escherichia coli* K12. When we proceed to rank this organism's proteome, half of the known drug targets appear at the top 10% of the ranking. The full distribution is displayed in Figure 4. It is interesting to note that this first half or top 10% correspond to essential proteins. Since essentiality is a binary score (*i.e.* genes may be essential or not), it divides the ranking in two sections as can be seen in the histogram (Figure 4). This clearly illustrates the difficulty of *a priori* ranking antibacterial targets, since even though essentiality is considered a very desirable property for any candidate target (Coates and Hu, 2007; Juhas *et al.*, 2012; Umland *et al.*, 2012), it only represents half of the known targets in this organism. Moreover, assessing gene essentiality is not a trivial task, given that *in vitro* results do not always correlate with gene essentiality determined *in vivo* (Umland *et al.*, 2012).

Because the equation that would unequivocally assign target scores to proteins is highly complex and full of unknowns, the pipeline presented here has been developed with the sole aim to assist the selection of prospective candidates, it is not meant to provide a final or complete list of antibacterial targets. Very often, it will not be used as a ranking tool but to retrieve target-relevant information for a specific protein and evaluate its pros and cons with respect to other potential candidates. The versatility of the tool, with dynamic features such as the assignation of relative weights at the user's criterion, and the availability of original information, such as predicted, functionally relevant ligand-binding sites, may prove valuable arguments for the microbiologist or medicinal chemist researching on new antibacterial targets.

Methods

Normalized sequence-similarity score (NS)

We used the BLAST program (Altschul *et al.*, 1997) with default parameters to scan complete proteomes. Since BLAST E-values may vary depending on the size of the queried database, we aligned all matched pairs and calculated their Smith-Waterman similarity score (Smith and Waterman, 1981). We ignored alignments with scores lower than 100, as previously described

(Aoki and Kanehisa, 2005).

Given that the Smith-Waterman similarity score is related to the size of the alignment, we divided the score by the length of the alignment to obtain a normalized sequence-similarity score (NS). The Smith-Waterman algorithm computes an optimal local alignment, meaning that the NS measure of similarity between two proteins is equivalent to the similarity between their most closely related pair of domains or regions.

Essentiality

Experimental information regarding gene essentiality is available for a few organisms at the database of essential genes (DEG) (Zhang and Zhang, 2008). If a particular strain was not available at DEG, we mapped query proteins to essential genes by using BLAST. For each annotated essential gene in a related strain, we scanned the proteome of interest and marked the best hit as an essential gene. Only E-values of 1e-10 or better were considered acceptable for this task. At the time of this writing, we were only able to gather large-scale essential gene information for: *Acinetobacter baumannii*, *Escherichia coli*, *Helicobacter pylori*, *Pseudomonas aeruginosa* and *Vibrio cholerae*.

Toxicity

An antibacterial drug acting on protein targets which are similar to human proteins may also bind these causing adverse effects and/or toxicity. We estimated the potential toxicity of each putative target proportional to the largest NS value obtained after pairwise alignment against the whole human proteome.

Isoforms and paralogs

If a given drug target presents multiple isoforms or paralogs ('variants'), the pathogen may readily develop resistance by functional substitution mechanisms. It is also possible that the drug may bind both the target and its variants, thus decreasing the antibiotic effect. To assess this parameter for each potential drug target, we counted the amount of variants present in the same proteome. We considered as variants of a protein all similar proteins with a NS value equal or larger than 2.

Evolutionary conservation among Gram-negative organisms

We defined a score to estimate the evolutionary conservation of potential targets across Gram-negative (GN) organisms as shown in Equation 1.

$$GNC_p = \sum_{i=1}^{i=n-1} \max(NS_p) \tag{1}$$

Where GNC_p is the Gram-negative conservation score for protein p, computed by adding the highest NS value ($\max(NS_p)$) obtained against each of the different GN species (i) in the data set, with n being the total number of GN species.

Conservation among strains

We estimated the evolutionary conservation of proteins among different strains of the same species using the following score:

$$SC_p = \frac{\sum_{j=1}^{j=m-1} \max(NS_p)}{m} \quad (2)$$

where SC_p is the strain conservation score for protein p, computed by adding the highest NS value ($\max(NS_p)$) obtained against each other strain (j) of the selected species in the data set, with m the total number of distinct strains of the particular species.

Scoring function and ranking of potential drug targets

Each of the different scores is normalized by the largest value obtained across the selected organisms. Normalized values are then multiplied by 100 to obtain percentages, *i.e.* final scores range between 0 and 100.

Each independent score has an associated weight, which can be negative or positive. These weighting values can be set by the user. However, default values are provided as follows. *A priori* negative features of a putative target (*i.e.* Toxicity and Paralogs) are given a default weight of -1, while positive features (*e.g.* Evolutionary conservation, Essentiality) have a corresponding default weight of 1.

For each protein in the selected data set, normalized scores are multiplied by their respective weights. The final score for each protein is obtained by summing up all weighted scores. Finally, all proteins in the selected data set are ranked according to their final score in terms of drug-target potential.

Comparative-genomics reference data set

Sequence data on Gram-negative (GN) organisms was gathered for a total of 770 fully sequenced GN proteomes covering 494 distinct species. GN bacteria species were identified at 'http://bacterialphylogeny.info/bacteria.html' and listed fully sequenced bacterial proteomes from 'http://www.uniprot.org/taxonomy' using the query string: '*bacteria AND complete:yes*'. A total of 770 bacterial strains were common to both listings. We downloaded sequence data from 'ftp://ftp.expasy.org/databases/complete_proteomes/fasta/bacteria/'.

Known drug targets and virulence factors

Each proteome of interest was compared by means of the BLAST program (Altschul *et al.*, 1997), with default parameters, against known drug targets available at DrugBank (Wishart *et al.*, 2008) and virulence factors available at VFDB (Yang *et al.*, 2008). Proteins showing a match with a BLAST E-value <= 1e-2 display a link to the related hits in the output table.

Three-dimensional homology models

An automated homology-modeling pipeline was implemented based on the program MODELLER v9.5 (Eswar *et al.*, 2008). Briefly, for a given protein sequence the system scans a

database of structural templates. It then proceeds to generate homology models using the best possible set of divers templates that display at least 30% sequence identity. Finally, the best resulting models are selected using a combination of DOPE and GA341 scores (Eswar *et al.*, 2008).

Acknowledgments

This work has been supported by funding under the Seventh Research Framework Programme of the European Union (ref. HEALTH-F3-2009-223101, AntiPathoGN) and the Spanish Ministry for Science and Innovation (BFU2010-17199). AP acknowledges the FPU Scholarship received from the Spanish Government. The authors acknowledge Manuel Alonso Tarajano for the initial compilation of the Gram-negative reference set.

Figures

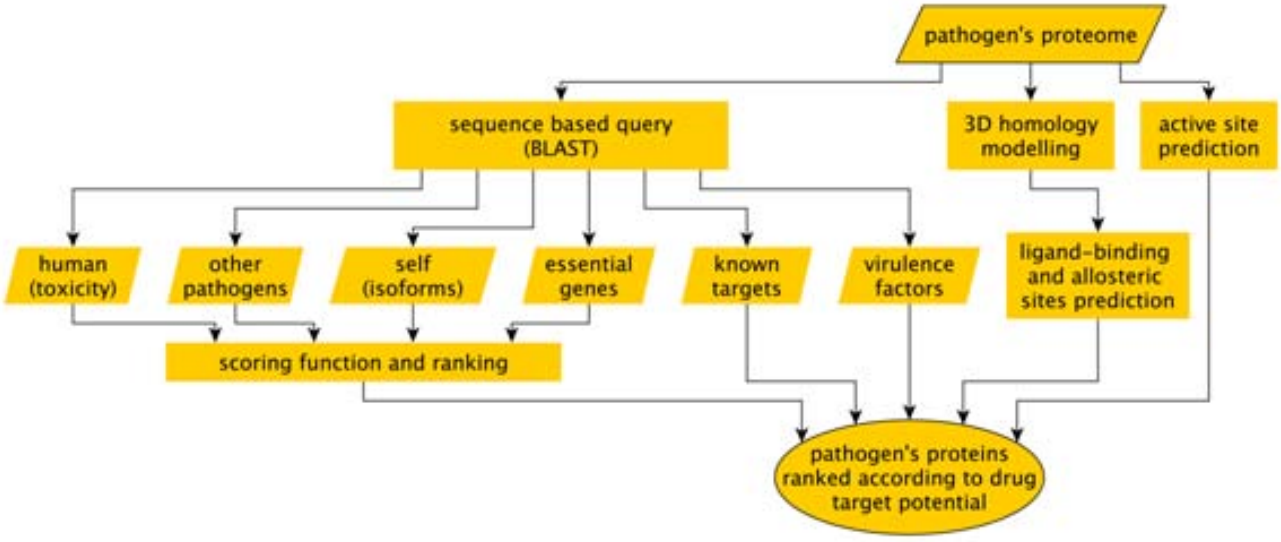

Figure 1: **Each genome of interest was subjected to the computational pipeline depicted in the flowchart.**

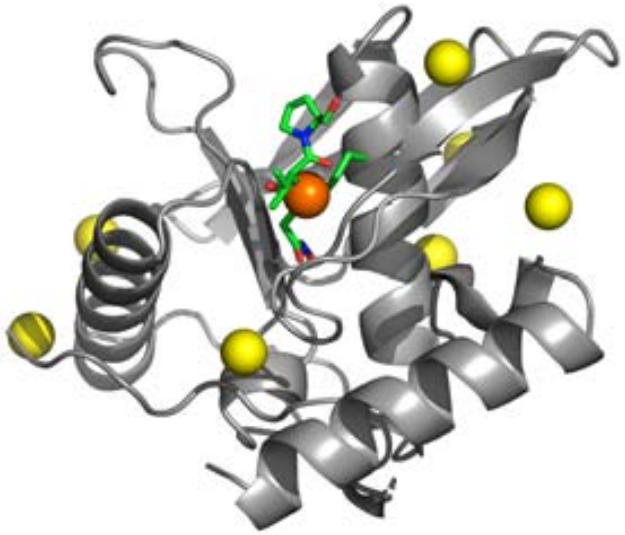

Figure 2: **Automatic modelling and structural pocket analysis performed on *Pseudomonas aeruginosa* deformylase [Swiss-Prot:Q9I7A8].** The spheres displayed on the homology model (based on template PDB:1N5N) represent putative ligand-binding sites as predicted by the automatic pocket analysis. The orange sphere marks the only cavity predicted to significantly affect overall protein flexibility. To illustrate the relevance of this prediction, we show the location of the antibiotic ligand (in 'sticks' representation) after superimposing the homology model to the known structure of the antibiotic-bound protein [PDB:1LRY] (RMSD 0.5 Å). The position of antibiotic Actinonin matches precisely the cavity marked by the procedure. The same cavity is also estimated to be very well conserved at the structural level (100% presence in the protein family).

Figure 3: **Automatic modeling and structural pocket analysis performed on *Escherichia coli* 3-oxoacyl-[acyl-carrier-protein] synthase 3 [Swiss-Prot:P0A6R0].** The spheres displayed on the homology model (based on template PDB:1UB7) represent putative ligand-binding sites as predicted by the automatic pocket analysis. The orange sphere marks the only cavity predicted to significantly affect overall protein flexibility. To illustrate the relevance of this prediction, we show the location of the inhibitor molecule (in 'sticks' representation) after superimposing the homology model to the known structure of the inhibitor-bound protein [PDB:1MZS] (RMSD 1.1 Å).

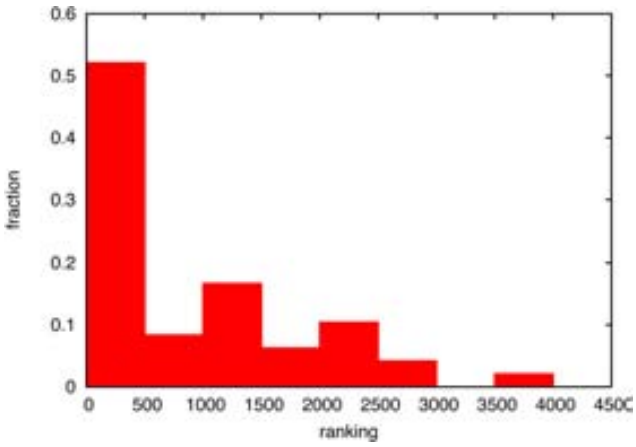

Figure 4: **The histogram displays the distribution of ranks for 48 known targets in *Escherichia coli* strain K12 within a ranking using default parameters.**

Bibliography

Akova,M., Daikos,G.L., Tzouvelekis,L. and Carmeli,Y. (2012) Interventional strategies and current clinical experience with carbapenemase-producing gram-negative bacteria. *Clin Microbiol Infect,* **18** (5), 439–448.

Altschul,S.F., Madden,T.L., Schaeffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res,* **25** (17), 3389–3402.

Aoki,K.F. and Kanehisa,M. (2005) Using the kegg database resource. *Curr Protoc Bioinformatics,* **Chapter 1**, Unit 1.12.

Arias,C.A. and Murray,B.E. (2009) Antibiotic-resistant bugs in the 21st century–a clinical superchallenge. *N Engl J Med,* **360** (5), 439–443.

Baquero,F., Alvarez-Ortega,C. and Martinez,J.L. (2009) Ecology and evolution of antibiotic resistance. *Environmental Microbiology Reports,* **1** (6), 469–476.

Bax,B.D., Chan,P.F., Eggleston,D.S., Fosberry,A., Gentry,D.R., Gorrec,F., Giordano,I., Hann,M.M., Hennessy,A., Hibbs,M., Huang,J., Jones,E., Jones,J., Brown,K.K., Lewis,C.J., May,E.W., Saunders,M.R., Singh,O., Spitzfaden,C.E., Shen,C., Shillings,A., Theobald,A.J., Wohlkonig,A., Pearson,N.D. and Gwynn,M.N. (2010) Type iia topoisomerase inhibition by a new class of antibacterial agents. *Nature,* **466** (7309), 935–940.

Bereket,W., Hemalatha,K., Getenet,B., Wondwossen,T., Solomon,A., Zeynudin,A. and Kannan,S. (2012) Update on bacterial nosocomial infections. *Eur Rev Med Pharmacol Sci,* **16** (8), 1039–1044.

Boucher,H.W., Talbot,G.H., Bradley,J.S., Edwards,J.E., Gilbert,D., Rice,L.B., Scheld,M., Spellberg,B. and Bartlett,J. (2009) Bad bugs, no drugs: no eskape! an update from the infectious diseases society of america. *Clin Infect Dis,* **48** (1), 1–12.

Brittnacher,M.J., Fong,C., Hayden,H.S., Jacobs,M.A., Radey,M. and Rohmer,L. (2011) Pgat: a multistrain analysis resource for microbial genomes. *Bioinformatics,* **27** (17), 2429–2430.

Chanumolu,S.K., Rout,C. and Chauhan,R.S. (2012) Unidrug-target: a computational tool to identify unique drug targets in pathogenic bacteria. *PLoS One,* **7** (3), e32833.

Coates,A.R.M. and Hu,Y. (2007) Novel approaches to developing new antibiotics for bacterial infections. *Br J Pharmacol,* **152** (8), 1147–1154.

Consortium,U. (2009) The universal protein resource (uniprot) 2009. *Nucleic Acids Res,* **37** (Database issue), D169–D174.

Council,E.A.S.A. (2007). Tackling antibacterial resistance in europe. Technical report EASAC.

Eswar,N., Eramian,D., Webb,B., Shen,M.Y. and Sali,A. (2008) Protein structure modeling with modeller. *Methods Mol Biol,* **426**, 145–159.

Fournier,P.E., Vallenet,D., Barbe,V., Audic,S., Ogata,H., Poirel,L., Richet,H., Robert,C., Mangenot,S., Abergel,C., Nordmann,P., Weissenbach,J., Raoult,D. and Claverie,J.M. (2006) Comparative genomics of multidrug resistance in acinetobacter baumannii. *PLoS Genet,* **2** (1), e7.

Guilloteau,J.P., Mathieu,M., Giglione,C., Blanc,V., Dupuy,A., Chevrier,M., Gil,P., Famechon,A., Meinnel,T. and Mikol,V. (2002) The crystal structures of four peptide deformylases bound to the antibiotic actinonin reveal two distinct types: a platform for the structure-based design of antibacterial agents. *J Mol Biol,* **320** (5), 951–962.

Huang,B. and Schroeder,M. (2006) Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct Biol,* **6**, 19.

Imming,P., Sinning,C. and Meyer,A. (2006) Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov,* **5** (10), 821–834.

Juhas,M., Stark,M., von Mering,C., Lumjiaktase,P., Crook,D.W., Valvano,M.A. and Eberl,L. (2012) High confidence prediction of essential genes in burkholderia cenocepacia. *PLoS One,* **7** (6), e40064.

Khandekar,S.S., Daines,R.A. and Lonsdale,J.T. (2003) Bacterial beta-ketoacyl-acyl carrier protein synthases as targets for antibacterial agents. *Curr Protein Pept Sci,* **4** (1), 21–29.

Kos,V.N., Desjardins,C.A., Griggs,A., Cerqueira,G., Tonder,A.V., Holden,M.T.G., Godfrey,P., Palmer,K.L., Bodi,K., Mongodin,E.F., Wortman,J., Feldgarden,M., Lawley,T., Gill,S.R., Haas,B.J., Birren,B. and Gilmore,M.S. (2012) Comparative genomics of vancomycin-resistant staphylococcus aureus strains and their positions within the clade most commonly associated with methicillin-resistant s. aureus hospital-acquired infection in the united states. *MBio,* **3** (3).

Laskowski,R.A., Rullmannn,J.A., MacArthur,M.W., Kaptein,R. and Thornton,J.M. (1996) Aqua and procheck-nmr: programs for checking the quality of protein structures solved by nmr. *J Biomol NMR,* **8** (4), 477–486.

Lee,D.S., Burd,H., Liu,J., Almaas,E., Wiest,O., Barabsi,A.L., Oltvai,Z.N. and Kapatral,V. (2009) Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple staphylococcus aureus genomes identify novel anti-microbial drug targets. *J Bacteriol,* **E**, Epub.

Lee,J.K., Lee,Y.S., Park,Y.K. and Kim,B.S. (2005) Mutations in the gyra and parc genes in ciprofloxacin-resistant clinical isolates of acinetobacter baumannii in korea. *Microbiol Immunol,* **49** (7), 647–653.

Mills,S.D. (2006) When will the genomics investment pay off for antibacterial discovery? *Biochem Pharmacol,* **71** (7), 1096–1102.

Mistry,J., Bateman,A. and Finn,R.D. (2007) Predicting active site residue annotations in the pfam database. *BMC Bioinformatics,* **8**, 298.

Nie,Z., Perretta,C., Lu,J., Su,Y., Margosiak,S., Gajiwala,K.S., Cortez,J., Nikulin,V., Yager,K.M., Appelt,K. and Chu,S. (2005) Structure-based design, synthesis, and study of potent inhibitors of beta-ketoacyl-acyl carrier protein synthase iii as potential antimicrobial agents. *J Med Chem,* **48** (5), 1596–1609.

Panjkovich,A. and Daura,X. (2010) Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC Struct Biol,* **10**, 9.

Panjkovich,A. and Daura,X. (2012) Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics,* **13** (1), 273.

Payne,D.J., Gwynn,M.N., Holmes,D.J. and Pompliano,D.L. (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov,* **6** (1), 29–40.

Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L., Poux,S., Redaschi,N., Xenarios,I., Bridge,A. and Consortium,U. (2013) Hamap in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Res,* **41** (Database issue), D584–D589.

Piras,C., Soggiu,A., Bonizzi,L., Gaviraghi,A., Deriu,F., Martino,L.D., Iovane,G., Amoresano,A. and Roncada,P. (2012) Comparative proteomics to evaluate multi drug resistance in escherichia coli. *Mol Biosyst,* **8** (4), 1060–1067.

Sakharkar,K.R., Sakharkar,M.K. and Chow,V.T.K. (2004) A novel genomics approach for the identification of drug targets in pathogens, with special reference to pseudomonas aeruginosa. *In Silico Biol,* **4** (3), 355–360.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J Mol Biol,* **147** (1), 195–197.

Souli,M., Galani,I. and Giamarellou,H. (2008) Emergence of extensively drug-resistant and pandrug-resistant gram-negative bacilli in europe. *Euro Surveill,* **13** (47).

Talbot,G.H., Bradley,J., Edwards,J.E., Gilbert,D., Scheld,M., Bartlett,J.G. and of the Infectious Diseases Society of America,A.A.T.F. (2006) Bad bugs need drugs: an update on the development pipeline from the antimicrobial availability task force of the infectious diseases society of america. *Clin Infect Dis,* **42** (5), 657–668.

Tally,F.P. and DeBruin,M.F. (2000) Development of daptomycin for gram-positive infections. *J Antimicrob Chemother,* **46** (4), 523–526.

Umland,T.C., Schultz,L.W., MacDonald,U., Beanan,J.M., Olson,R. and Russo,T.A. (2012) In vivo-validated essential genes identified in acinetobacter baumannii by using human ascites overlap poorly with essential genes detected on laboratory media. *MBio,* **3** (4).

van Hal,S.J. and Paterson,D.L. (2011) New gram-positive antibiotics: better than vancomycin? *Curr Opin Infect Dis,* **24** (6), 515–520.

White,T.A. and Kell,D.B. (2004) Comparative genomic assessment of novel broad-spectrum targets for antibacterial drugs. *Comp Funct Genomics,* **5** (4), 304–327.

Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2008) Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res,* **36** (Database issue), D901–D906.

Yang,J., Chen,L., Sun,L., Yu,J. and Jin,Q. (2008) Vfdb 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res,* **36** (Database issue), D539–D542.

Zhang,C.T. and Zhang,R. (2008) Gene essentiality analysis based on deg, a database of essential genes. *Methods Mol Biol,* **416**, 391–400.

Zhang,H.J., Li,Z.L. and Zhu,H.L. (2012) Advances in the research of beta-ketoacyl-acp synthase iii (fabh) inhibitors. *Curr Med Chem,* **19** (8), 1225–1237.