



Development of tools for in silico drug discovery

Adrià Cereto-Massagué

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

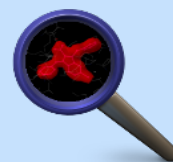
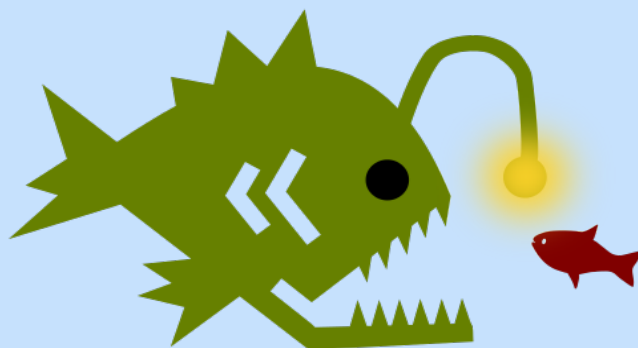
WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



UNIVERSITAT
ROVIRA I VIRGILI

Development of tools for *in silico* drug discovery

ADRIÀ CERETO-MASSAGUÉ



TESI DOCTORAL – TESIS DOCTORAL – DOCTORAL THESIS
2017

UNIVERSITAT ROVIRA I VIRGILI

Development of tools for in silico drug discovery

Adrià Cereto-Massagué

UNIVERSITAT ROVIRA I VIRGILI

Development of tools for in silico drug discovery

Adrià Cereto-Massagué

UNIVERSITAT ROVIRA I VIRGILI

Development of tools for in silico drug discovery

Adrià Cereto-Massagué

Development of tools for *in silico* drug discovery

Adrià Cereto-Massagué

Doctoral thesis

Supervised by Dr. Santiago Garcia-Vallvé and Dr.
Gerard Pujadas
Department of Biochemistry and Biotechnology

2017



UNIVERSITAT ROVIRA i VIRGILI

UNIVERSITAT ROVIRA I VIRGILI

Development of tools for in silico drug discovery

Adrià Cereto-Massagué



UNIVERSITAT
ROVIRA I VIRGILI

FEM CONSTAR que aquest treball, titulat “Development of tools for *in silico* drug discovery”, que presenta n’Adrià Cereto i Massagué per a l’obtenció del títol de Doctor amb menció internacional, ha estat realitzat sota la nostra direcció al Departament de Bioquímica i Biotecnologia d’aquesta universitat.

HACEMOS CONSTAR que el presente trabajo, titulado “Development of tools for *in silico* drug discovery”, que presenta Adrià Cereto i Massagué para la obtención del título de Doctor con mención internacional, ha sido realizado bajo nuestra dirección en el Departamento de Bioquímica y Biotecnología de esta universidad

WE STATE that the present study, entitled “Development of tools for *in silico* drug discovery”, presented by Adrià Cereto-Massagué for the award of the degree of Doctor with international mention, has been carried out under our supervision at the Department of Biochemistry and Biotechnology of this university.

Tarragona, 2-6-2017

Els directors de la tesi doctoral
Los directores de la tesis doctoral
Doctoral Thesis Supervisors

Santiago Garcia-Vallvé

Gerard Pujadas

UNIVERSITAT ROVIRA I VIRGILI

Development of tools for in silico drug discovery

Adrià Cereto-Massagué

Index

| | |
|--|----|
| Agraiments..... | 1 |
| Introduction..... | 3 |
| Objectives..... | 8 |
| Molecular Fingerprint Similarity Search in Virtual Screening..... | 9 |
| Abstract..... | 10 |
| 1. Introduction..... | 11 |
| 2. Methods for Molecular Fingerprints..... | 11 |
| 2.1 Types of Molecular Fingerprint..... | 12 |
| 2.2 Software for fingerprint-based virtual screening..... | 17 |
| 2.3 Online tools for fingerprint-based virtual screening..... | 21 |
| 3. Usual fingerprint-based virtual screening scenarios..... | 23 |
| 4. Comparing Fingerprint Similarity Search with other Virtual Screening Methods..... | 24 |
| 5. Conclusion..... | 27 |
| 6. Acknowledgements..... | 28 |
| 7. References..... | 29 |
| 8. Appendix..... | 38 |
| Figures:..... | 39 |
| DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets..... | 41 |
| Abstract..... | 41 |
| Introduction..... | 42 |
| Program overview..... | 43 |
| Implementation and system requirements..... | 45 |
| Acknowledgements..... | 46 |
| References..... | 46 |
| Molecular weight-based decoys: a simple decoy set finding alternative for fingerprint similarity approaches..... | 49 |
| Abbreviations..... | 49 |
| Abstract..... | 50 |
| Introduction..... | 51 |
| Computational methods..... | 53 |
| Results and discussion..... | 55 |
| Conclusion..... | 58 |
| Tables..... | 60 |
| Author information..... | 64 |
| Acknowledgements..... | 65 |
| References..... | 65 |
| Figures..... | 68 |
| The Good, the Bad and the Dubious. VHELIBS, a Validation Helper for Ligands and Binding Sites..... | 75 |
| Abstract..... | 75 |
| Keywords..... | 76 |

| | |
|--|-----|
| Background..... | 76 |
| Implementation..... | 78 |
| Description of the algorithm..... | 79 |
| Key features of VHELIBS..... | 82 |
| PDB_REDO changes to support VHELIBS..... | 83 |
| Results and Discussion..... | 83 |
| Conclusions..... | 86 |
| Availability and Requirements..... | 87 |
| List of abbreviations..... | 87 |
| Competing interests..... | 87 |
| Author's contributions..... | 87 |
| Acknowledgements..... | 88 |
| References..... | 88 |
| Figures..... | 92 |
| Tables..... | 96 |
| Tools for <i>In Silico</i> Target Fishing..... | 101 |
| Abstract..... | 101 |
| Highlights..... | 103 |
| Keywords..... | 103 |
| 1. Introduction..... | 104 |
| 2. Computational methods for target fishing..... | 105 |
| 2.1 Molecular similarity methods..... | 105 |
| 2.2 Data mining and machine learning methods..... | 108 |
| 2.3 Protein structure-based methods..... | 109 |
| 2.4 Methods based on analysis of bioactivity spectra..... | 109 |
| 3. Validation of the methods..... | 110 |
| 4. Examples of target predictions..... | 112 |
| 5. Conclusions..... | 114 |
| Acknowledgements..... | 115 |
| References..... | 115 |
| Figures..... | 126 |
| Anglerfish: a webserver for quantitative prediction of ligand bioactivity..... | 129 |
| Keywords..... | 129 |
| Implementation and Methods..... | 132 |
| Molecule Database..... | 132 |
| Interface..... | 134 |
| Validation..... | 140 |
| Results and Discussion..... | 143 |
| Conclusions..... | 146 |
| Author Information..... | 147 |
| References..... | 148 |
| Conclusions..... | 153 |
| References..... | 155 |

Agraïments

A la Meritxell, per ser-hi sempre.

I a tots aquells que m'han acompanyat durant aquests anys.

Introduction

Drug discovery is one of the most important research fields for the future of health care and the advancement of most biological sciences. The computational advances during the last decades have enabled the emergence of cheminformatics (or *chemoinformatics*^{1,2}) as a new discipline essential to the drug discovery process. Even though cheminformatics has expanded beyond its original drug discovery scope, that is still the main focus of cheminformatics, with packages and tools available for all the parts of the drug discovery process that can be performed *in silico*.

However, there is always room for improvement, and, as a relatively young discipline, there are still plenty of gaps that need to be filled and problems that need to be solved. The aim of this thesis was to identify some of those gaps and to fill them by developing new tools that could prove useful for others and to make easier and improve the overall quality of research with cheminformatics.

Virtual screening is a cheminformatics method that consists of screening large small-molecule databases for bioactive molecules^{3,4}. This enables the researcher to avoid the cost of experimentally testing hundreds or thousands of compounds by reducing the number of candidate molecules to be tested to manageable numbers. There are several different approaches to it, but they are ultimately all based on either the biological target of the prospective drug, or its known bioactive molecules^{3,4}. However, in order to assess the reliability of each these methods, regardless of their approach, they all need to be validated. This can be done by putting together known bioactive molecules and known inactive

molecules⁴, and seeing how well does the virtual screening method identify the bioactive molecules from within the inactive molecules. There is a problem, though: in most cases, there is not enough available data of inactive molecules, as negative results are less likely to be reported. So a solution is to build a library of decoy molecules, that is, molecules that are not trivially different from the bioactive compounds, but not too similar to them, either, as that could lead to the inclusion of unknown actually bioactive molecules (false negatives). In order to minimize the impact of false negatives, decoy libraries are most often built to be much bigger than known bioactive molecule libraries, with a ratio of several decoy molecules for each bioactive compound. This way, if the virtual screening method manages to successfully pick the bioactive molecules out of the majority of decoy molecules (high enrichment), its reliability gets validated⁵. At the time when this thesis was started, there were some decoy library databases available, like the Directory of Useful Decoys⁵, which only covered a certain set of targets. There weren't any tools available for building decoy libraries for any arbitrary set of bioactive compounds, and in order to provide that and fill the gap, DecoyFinder, an easy to use graphical application, was developed. It was later updated after some research into decoy library building and their performance when used for 2D similarity approaches.

Cheminformatics relies a lot on the data available in publicly accessible repositories like ChEMBL or the Protein Data Bank (PDB)⁶. The PDB is very useful because it provides the 3D structures of thousands of protein-ligand complexes and, therefore, provides information on how certain ligands bind and interact with their targets, which is very valuable for some widely used virtual screening approaches. One such approach is molecular docking, where the binding position and energy of a molecule with a target is simulated and used to

predict whether it will bind as a ligand or not⁷. Another approach consist in using the spatial location of the intermolecular interactions between a ligand and its target (what is called a *pharmacophore*) to find other small molecules that can make the same kind of interactions with the target⁸. For such approaches, it is of the utmost importance that the data available on the PDB for the 3D structure of the ligand and its binding site are reliable. However, this is not the case for all PDB entries, as can be seen by inspecting the electron density maps of the relevant residues and atoms for each structure, in some cases even revealing ligands with little evidence of actually being there or residues that may be completely wrong. However, this is far from intuitive or easy to do for non-crystallographers, and thus, VHELIBS was developed as a tool to easily and intuitively inspect and identify reliable PDB structures based on the goodness of fitting between ligands and binding sites and their corresponding electron density map, also leveraging the often more accurate structures from PDB_REDO⁹.

While virtual screening aims to find new bioactive molecules for certain targets, the opposite approach is also used: starting from a given molecule, to search for a biological target for which it presents previously undocumented bioactivity. This reverse screening is known as *in silico* or computational target fishing¹⁰ or reverse pharmacognosy¹¹, and it is specially useful for drug repurposing or repositioning¹². Drug repurposing consists on finding new biological targets for already approved drugs. This can potentially save hundreds of millions of dollars and more than a decade of research when compared to a novel drug for the same target, since its safety and suitability will have already been studied¹³. Target fishing approaches can also uncover polypharmacology for a given molecule (when it presents bioactivity for several

targets), which is important since known drugs have an average of six molecular targets on which they exhibit activity¹⁴. They can also help detect potentially toxic side effects and thus help develop less toxic drugs¹⁵. When this thesis began, there were no freely available target fishing platforms, but some have been developed during the years, as can be seen in *Tools for in silico target fishing* review. However, they are qualitative in the nature of their activity prediction, and thus we developed a freely accessible target fishing web service implementing a novel method which provides the first quantitative activity prediction: Anglerfish.

- (1) Gasteiger, J.; Engel, T. *Chemoinformatics*; Gasteiger, J., Engel, T., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, FRG, 2003.
- (2) Brown, F. K. In *Annual Reports in Medicinal Chemistry*; 1998; Vol. 33, pp 375–384.
- (3) Rollinger, J. M.; Stuppner, H.; Langer, T. *Prog. Drug Res.* **2008**, *65*, 211, 213–249.
- (4) Rester, U. *Curr. Opin. Drug Discov. Devel.* **2008**, *11* (4), 559–568.
- (5) Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.
- (6) Park, W.S., Kang, S.K., Jun, M.A., Shin, M.S., Kim, K.Y., Rhee, S.D., Bae, M.A., Kim, M.S., Kim, K.R., Kang, N.S., Yoo, S.E., Lee, J.O., Song, D.H., Silinski, P., Schneider, S.E., Ahn, J.H., Kim, S. S. RCSB Protein Data Bank - RCSB PDB <http://www.rcsb.org/pdb/explore/explore.do?structureId=3Q8W> (accessed Mar 7, 2013).
- (7) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. *P. J. Med. Chem.* **2005**, *48* (4), 962–976.

- (8) Koes, D. *Comput. Drug Discov.* **2016**.
- (9) Joosten, R. P.; Long, F.; Murshudov, G. N.; Perrakis, A. *IUCrJ* **2014**, *1* (4), 213–220.
- (10) Liu, X.; Xu, Y.; Li, S.; Wang, Y.; Peng, J.; Luo, C.; Luo, X.; Zheng, M.; Chen, K.; Jiang, H. *J. Cheminform.* **2014**, *6* (1), 33.
- (11) Blondeau, S.; Do, Q. T.; Scior, T.; Bernard, P.; Morin-Allory, L. *Curr. Pharm. Des.* **2010**, *16* (15), 1682–1696.
- (12) Ashburn, T. T.; Thor, K. B. *Nat. Rev. Drug Discov.* **2004**, *3* (8), 673–683.
- (13) Abrutyn, E. *Infect. Dis. Clin. North Am.* **1989**, *3* (3), 653–664.
- (14) Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. *Mol. Biosyst.* **2009**, *5* (9), 1051.
- (15) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. *Nature* **2012**.

Objectives

- I.** To create a freely available tool with a graphical interface to facilitate the validation of virtual screening approaches by making decoy molecule library building easier and more accessible.

- II.** To develop a freely available software to enable non-experts to intuitively and visually assess the quality and reliability of the 3D crystallographic structures of ligands and binding sites in the Protein Data Bank

- III.** To develop a publicly available target fishing service with quantitative bioactivity prediction.

Molecular Fingerprint Similarity Search in Virtual Screening

*Adrià Cereto-Massagué¹, María José Ojeda¹, Cristina Valls¹, Miquel Mulero¹,
Santiago Garcia-Vallvé^{1,2} and Gerard Pujadas^{1,2*}*

¹Group of Cheminformatics & Nutrition. Biochemistry and Biotechnology Department, Universitat Rovira i Virgili (URV), Campus de Sescelades, N4 Building, 43007 Tarragona, Catalonia, Spain

² Centre Tecnològic de Nutrició i Salut (CTNS), TECNIO, CEICS, Avinguda Universitat, 1, 43204, Reus, Spain

* Corresponding author: gerard.pujadas@urv.cat

Abstract

Molecular fingerprints have been used for a long time now in drug discovery and virtual screening. Their ease of use (requiring little to no configuration) and the speed at which substructure and similarity searches can be performed with them –paired with a virtual screening performance similar to other more complex methods– is the reason for their popularity. However, there are many types of fingerprints, each representing a different aspect of the molecule, which can greatly affect search performance. This review focuses on commonly used fingerprint algorithms, their usage in virtual screening, and the software packages and online tools that provide these algorithms.

Keywords: Fingerprints, virtual screening, similarity search, data fusion, comparison

1. Introduction

Computational advances during the past two decades have enabled the extensive use of virtual screening for drug discovery [1]. Virtual screening is an *in silico* method that consists of screening large small-molecule databases for bioactive molecules. This enables the researcher to avoid the cost of experimentally testing hundreds or thousands of compounds by reducing the number of candidate molecules to be tested to manageable numbers.

The screening can be conducted using several methods or their combination, which can be classified as structure-based methods (which are based on matching the compounds to a target binding site, the most common of these approaches being protein-ligand docking) or ligand-based methods (which involves retrieving those compounds from the database that are similar in some ways to known active molecules and vary greatly depending on the molecular features taken into account for similarity assessment). The main ligand-based approaches involve the use of pharmacophores (abstractions of the features needed for the molecule to be active) [2], shape-based similarity [3], fingerprint similarity, and also machine learning using molecular properties and data from any of the former approaches [4].

Fingerprint-based similarity searching also sees some use in other fields besides virtual screening and drug discovery, such as flavor chemistry [5].

2. Methods for Molecular Fingerprints

Similarity in itself is subjective and can be measured and their results interpreted in several ways [6–8]. One of the most important problems encountered when trying to measure the similarity between two compounds is

the complexity of the task, which depends on the complexity of the molecular representation used. In order to make the comparison between molecular representations computationally easier, some level of simplification or abstraction is required. The most commonly used of these abstractions are molecular fingerprints, which involve turning the molecule into a sequence of bits that can then be easily compared between molecules.

This comparison must then be expressed in a way that can be quantified. There are many ways to assess the similarity between two vectors, the most common overall being Euclidean distance. But for molecular fingerprints, the industry standard is the Tanimoto coefficient, which consists of the number of common bits set to 1 in both fingerprints divided by the total number of bits set to 1 between both fingerprints. This means that it will always have a value between 1 and 0, regardless of length of the fingerprint, which causes it to lose representativity the longer the fingerprints are. This also means that how actually similar two fingerprints are with a given Tanimoto coefficient value will greatly depend on the type of fingerprint used, which makes impossible the selection of an universal cutoff criterion from which two fingerprints could be deemed similar or dissimilar. However, the performance of molecular fingerprints could be improved by combining it with the use of other similarity coefficients [9]. Several similarity and distance metrics which have been used with fingerprints can be seen in Table 1.

2.1 Types of Molecular Fingerprint

There are several types of molecular fingerprints depending on the method by which the molecular representation is transformed into a bit string. Most of them use only the 2D molecular graph and are thus called 2D fingerprints, but

some are capable of storing 3D information, most notably pharmacophore fingerprints. The main approaches are substructure keys-based fingerprints, topological or path-based fingerprints, and circular fingerprints.

- **Substructure keys-based fingerprints** set the bits of the bit string depending on the presence in the compound of certain substructures or features from a given list of structural keys. This usually means that these fingerprints are most useful when used with molecules that are likely to be mostly covered by the given structural keys, but not so much when the molecules are unlikely to contain the structural keys, as their features would not be represented. Their number of bits is determined by the number of structural keys, and each bit relates to presence or absence of a single given feature in the molecule (Figure 1), which does not happen with other (hashed) types of fingerprints. Some of the most commonly used substructure keys-based fingerprints are:
 - MACCS [10,11]: It comes in two variants, one with 960 and the other with 166 structural keys based on SMARTS patterns. The shorter one is the most commonly used, as it is relatively small in length (only 166 bits) but covers most of the interesting chemical features for drug discovery and virtual screening. Additionally several software packages are able to calculate it, which is not true for the longer version.
 - PubChem fingerprint [12]: this fingerprint, with 881 structural keys covers a wide range of different substructures and features. It is the fingerprint used by PubChem for similarity searching and neighboring. Other than PubChem's own code, it is also

implemented in ChemFP [13] (although deemed “experimental”) and in CDK [14,15].

- BCI fingerprints [16]: BCI fingerprints can be generated using different amounts of bits and can be modified by the user in several ways, but the standard substructure dictionary includes 1052 keys [17]. BCI fingerprints are only available in BCI toolkits.
- TGD [18] and TGT fingerprints: These are two-point and three-point pharmacophoric fingerprints calculated from a 2D molecular graph, consisting, respectively of 735 and 13824 bits. TGD encodes atom-pair descriptors using seven-atom features and distances up to 15 bonds [17,18]. TGT encodes triplets of four-atom features using three graph distances divided into six distance ranges [17]. They are both available in MOE software package [19].
- **Topological or path-based fingerprints** work by analyzing all the fragments of the molecule following a (usually linear) path up to a certain number of bonds, and then hashing every one of these paths to create the fingerprint (Figure 2). This means that any molecule can produce a meaningful fingerprint, and its length can be adjusted. They can also be used for fast substructure searching and filtering. These are hashed fingerprints, which means that a single bit cannot be traced back to a given feature. A given bit may be set by more than one different feature, which is called “bit collision”. The Daylight fingerprint [20]: is the most prominent of these types of fingerprints. They consist of up to 2048 bits and encode all possible connectivity pathways through a molecule up to a given length. Most software packages implement these fingerprints or fingerprints based on them, which can sometimes reach

higher number of bits or use non-linear connectivity paths, such as OpenEye's Tree fingerprints [21]

- **Circular fingerprints** are also hashed topological fingerprints, but they are different in that instead of looking for paths in the molecule, the environment of each atom up to a determined radius is recorded. They are therefore not suitable for substructure queries (as the same fragment may have different environments) but are widely used for full structure similarity searching.
 - Molprint2D [22,23]: Molprint2D encodes the atom environments of each atom of the molecular connectivity table, which are represented by strings of varying size. This fingerprint is available in several software packages, such as OpenBabel [24] and jCompoundMapper [25].
 - ECFP: The *de facto* standard circular fingerprints are the Extended-Connectivity Fingerprints (ECFPs), based on the Morgan algorithm [26], which were specifically designed for their use in structure-activity modeling [27]. They represent circular atom neighborhoods and produce fingerprints of variable length. They are most commonly used with a diameter of 4 and referred to as ECFP4. A diameter of 6 (ECFP6) is also commonly used, although some benchmarks have shown small performance differences between the two [28]. Additionally, there is a variation that keeps track of the frequency counts of the ECFP features, recording each identifier as many times as it appears in the molecule instead of only once. This variation is often denoted as ECFC. Notable software programs that provide these fingerprints are Pipeline Pilot [29], Chemaxon's

JChem [30], the CDK [14] and the RDKit [31] (referred to as “Morgan fingerprints”).

- FCFP (Functional-Class Fingerprints): FCFP are a variation of ECFP, which are further abstracted in that instead of indexing a particular atom in the environment, they index that atom's role. So, different atoms or groups with the same or similar function are not distinguished by the fingerprint. This enables them to be used as pharmacophoric fingerprints. There is also a FCFC variation, akin to the ECFC variation to the ECFP. All major software packages supporting ECFP fingerprints also support these variations.
- There are also some hybrid fingerprints that combine the same bits string bits set using different approaches. Some commonly used fingerprints that fall into this category are the following:
 - UNITY 2D [32]: This is a 988-bit long fingerprint based both on structural keys and connectivity path fragments.
 - MP-MFP [33]: MP-MFP is a 171-bit fingerprint with 110 bits set from structural keys and 61 bits set from property descriptors.
- Pharmacophore fingerprints are also commonly used. A pharmacophore represents the relevant features and interactions needed for a molecule to be active against a given target. Pharmacophoric fingerprints usually encode the information for the features from a list that a molecule presents, in a similar way to substructure-key based fingerprints, but taking into account the distance between these features, usually

classifying it using a list of distance ranges. This way 3D information can be encoded into the fingerprint [34].

- Lastly, there are also other types of fingerprints that try totally different approaches. For example, LINGO [35] and SMIfp [36] are fingerprints that are text-based and are calculated based on the canonical SMILES [37] of the molecule. Protein-ligand interaction fingerprints (PLIF), as their name suggests, encode information on protein-ligand interactions, such as hydrogen bonds, ionic interactions and surface contacts with their residue of origin [19]. Structural Interaction Fingerprint (SIFt) is also one of these fingerprints [38].

In general, fingerprints with longer bit strings have been found to perform better at similarity searching, because of increased amount of stored information (due to a reduction of bit collision for hashed fingerprints) [39].

2.2 Software for fingerprint-based virtual screening

There are many software packages that can be used for fingerprint-based virtual screening, from whole drug discovery suites including fingerprint functionality to software libraries or tools centered specifically in dealing with fingerprints and similarity searching. Each software package supports a different set of fingerprints, and most of them implement fingerprints not present in any other package. However, the most commonly used fingerprinting algorithms can be found in most software packages. Here is a list of the main software packages used when doing ligand-based virtual screening with fingerprint similarity, in no specific order:

- OEChem TK: This OpenEye toolkit [21] is able to produce 166-bit MACCS, LINGO, Circular, Path (Daylight-like) and Tree (Daylight-like with non-linear, “tree” fragments) fingerprints. It has interfaces to C++, Java, Python, and C#.
- JChem from ChemAxon [30]: This is a java library that provides access to several hashed fingerprints, ECFP fingerprints with all their variants (ECFC, FCFP, FCFC), and pharmacophoric fingerprints. ChemAxon also provides packages for .NET and is usable in Python through cinfony [40].
- OpenBabel [24,41] This is a free and open-source cheminformatics toolkit, which implements MOLPRINT2D, 166-bit MACCS, a Daylight-like fingerprint (FP2), and 2 structural key fingerprints with 55 (FP3) and 307 bits. It can be used from C++, Java, Python, C#, and Perl.
- RDKit [31] RDKit is also a free and open-source cheminformatics toolkit that provides access to several fingerprints: 166-bit MACCS, “Topological” (Daylight-like), “Atom pairs” (based on the atomic environments and shortest path separations of every atom pair in the molecule [42]), “Morgan” (ECFP and its variations), “Torsion” (based on the topological torsion descriptor [43]), and “Layered” (an experimental topological fingerprint intended to make fingerprinting queries more straightforward). It is usable from C++, Python, Java, and C#.
- CDK [14,15,44], CDK is another free and open-source toolkit, which features several fingerprints, the most notable being ECFP, LINGO, Daylight-like fingerprint, 166-MACCS, PubChem, and other structural

keys fingerprints such as E-State [45] and Klekota-Roth [46]. It is a Java library but can be used in regular Python through cinfony [40].

- **Indigo:** Indigo, yet another free and open-source cheminformatics toolkit, offers several hashed fingerprints and their combination [47]. It can be used from C++, Java, Python and C#.
- **Cinfony [40,48]:** Cinfony is not a toolkit in itself and does not implement any fingerprint, but it gives the user access to several toolkits (OpenBabel, RDKit, CDK, JChem, and Indigo) through a common API in Python and to some extent in Jython (JVM) and IronPython (.NET).
- **ChemFP [13]:** ChemFP is a tool that can be used as a back-end database with either OpenBabel, RDKit or OEChem, thus supporting most of their fingerprints, and implementing on top of that a 166-bit MACCS and a PubChem-like fingerprint. But what is special about Chemfp is its ability to store the fingerprints in a standard file format (FPS) and then to perform high-speed Tanimoto similarity searches. It provides a Python library and command-line tools.
- **Canvas:** Canvas from Schrödinger offers MACCS, customizable SMARTS-based keys fingerprints, and seven types of hashed fingerprints, including MOLPRINT2D, ECFP, and linear (Daylight-like), as well as fingerprints derived from pharmacophore models [39,49,50].
- **Molecular Operating Environment (MOE):** MOE implements 2 (TGD), 3 (TGT), and 4-point pharmacophore fingerprints in 2D/3D, MACCS keys, and EigenSpectrum shape fingerprints among others [19].

- jCompoundMapper [25,51]: This is an open-source command-line tool and a library for chemical fingerprints, featuring support for many fingerprint types, including MOLPRINT2D, atom pairs, and pharmacophore fingerprints among others. It also provides several machine learning tools and uses CDK.
- Pipeline Pilot from Accelrys [29]: This is an authoring tool with a visual and dataflow authoring language. It can calculate a wide variety of fingerprints, including both MACCS versions, ECFP, and its variants.
- SYBYL-X Suite from Tripos [32]: SYBYL-X is a molecular modeling suite that includes the UNITY 2D fingerprints for similarity searches.
- DecoyFinder [52,53]: DecoyFinder is a graphical tool that helps find decoy sets for virtual screening validation. It uses MACCS fingerprints and molecular descriptors to find the decoy molecules.
- FLAP [54] (Fingerprints for Ligands and Proteins): FLAP is a tool that provides a common reference framework for comparing molecules using GRID Molecular Interaction Fields (MIFs). The fingerprints are characterized by quadruplets of pharmacophoric features and can be used for ligand-ligand, ligand-receptor, and receptor-receptor comparison.
- MayaChemTools is a free collection of Perl scripts, modules and classes to support day-to-day computational discovery needs [55]. It can compute several, molecular fingerprints, including ECFP, MACCS, path-based fingerprints and many more. It can also be directly used for similarity searching with fingerprints.

2.3 Online tools for fingerprint-based virtual screening

In comparison to the large number of software packages offering fingerprint functionality, the number of online services doing so is far lower, mostly consisting of databases that include a similarity searching option using some fingerprint. A brief enumeration of the most interesting services is as follows:

- PubChem [56]: PubChem provides a fast chemical structure similarity search tool. Any small molecule may be used as query, and a Tanimoto coefficient threshold can be chosen above which molecules will be deemed similar enough. The fingerprint used for this similarity searches is the PubChem fingerprint [12].
- ChemSpider [57–59]: ChemSpider also supports similarity searching with Tanimoto (and other metrics) thresholds. It uses a fingerprint calculated by GGA's BINGO database cartridge, which uses the Indigo toolkit [49].
- The ZINC database [60–62]: This database also supports similarity search. The fingerprint used is the path-based ChemAxon fingerprint from JChem [30,61]. It uses the same fingerprint for the generation of clusters with molecules of up to a given similarity cutoff, which produces clusters with guaranteed molecular diversity and chemical space coverage.
- Multi-Fingerprint Browser for ZINC [63,64]: This is a tool that enables rapid identification of close analogs among commercially available compounds in the ZINC database [60]. The browser retrieves nearest neighbors in multi-dimensional chemical spaces defined by four different fingerprints (fingerprint = a vector composed of several

numerical descriptors of molecular structure and properties), each of which represents relevant structural and pharmacophoric features in a different way: sFP (substructure fingerprint), ECFP4 (Extended connectivity fingerprint), MQN (Molecular Quantum Numbers), and SMIfp (SMILES fingerprint). Distances are calculated using the city-block distance (CBD; see Table 1), a similarity measure which, according to Awale et al. [63], performs as well as Tanimoto similarity.

3. Usual fingerprint-based virtual screening scenarios

To conduct a virtual screening based on fingerprint similarity, the following things are needed:

- At least one known active molecule, which will be the reference molecule(s).
- A molecular database with potential actives.
- Software capable of generating and comparing fingerprints.

Once the reference molecules are chosen, the next step would be to choose the most appropriate fingerprint. The choice is usually limited by the available options in the software being used. The most appropriate option would also depend mostly on the reference molecules, as a fingerprint should be able to properly represent the reference molecules (which is generally not a concern for hashed fingerprints). It should also be taken into account whether the database and the available fingerprints account for stereochemistry, tautomeric forms, and conformation of both the reference molecules and the molecules in the database to be screened. Stereochemistry-aware methods should be preferably used to screen equally stereochemistry-aware databases. The presence of conformations enables the use of fingerprints depending on them[34]. Tautomerism of the studied molecules should also be taken into account because different tautomers of the same molecule could have substantially different fingerprints.

With the chosen algorithm, fingerprints would be calculated for every molecule and reference in the database, and then the similarity coefficient is calculated between the reference molecule and every other molecule. After this, the

molecules can be ranked in descending order using the similarity coefficient. The top molecules of the rank would be expected to exhibit a similar activity as the reference molecule.

4. Comparing Fingerprint Similarity Search with other Virtual Screening Methods

In a comparison by Tresadern et al [65], ECFP6 fingerprints were compared to several other virtual screening methods: feature trees, topomers, ROCS shape Tanimoto, EON electrostatic Tanimoto, OpenEye ComboScore (a combination of shape Tanimoto and color-score), and Cresset-Fieldscreen. All of these, other than those that feature trees, are 3D methods and require substantially more computation time than fingerprints. The results were as expected: the ECFP6 fingerprint was the weakest performing method with 3 out of the 4 queries, although it exhibited one of the highest performances with the remaining query. However, the 3 queries, where the fingerprint was outperformed, all showed very similar performances for all the methods, which may imply that the performance of the methods depends on the selected queries.

In a different comparison, by McGaughey et al [66], the Daylight fingerprint was put to test against many other virtual screening methods, including protein-ligand docking. The Daylight fingerprint outperformed most of the other methods. The authors conclude that “as measured by EF, the 2D similarity methods (TOPOSIM, Daylight) perform well at lead-hopping when applied to a diverse database.[...] One may ask how it is possible for 2D similarity methods to perform nearly as well as 3D methods at lead hopping.” They also noted how sensitive the performance is in Daylight fingerprints regarding path length, and

that the default settings (minimum path length of 0 and maximum of 7) is too easy to outperform making them poor standards for 2D similarity.

In yet another comparison [67], several fingerprints (OpenBabel FP2, BCI, MACCS, Daylight and MOLPRINT2D) were compared against 3D molecular shape-based methods (ESHAPE3D, ROCS, PARAFIT, SHAEP and USR). Given the results, the authors state that “Overall, we find that the 2D fingerprint-based methods give better Virtual Screening performance than the 3D shape-based approaches for many of the DUD targets”. This shows how 3D methods do not always outperform simple fingerprint similarity search.

However, when comparing fingerprint similarity searching to other virtual screening approaches, the use of fingerprints has several advantages:

- It requires minimal setup and configuration. Some fingerprints can be fine-tuned in several ways, but it will still require a lot less work than creating pharmacophores or selecting and preparing a binding site for a protein-ligand docking.
- Most of the commonly used fingerprints are calculated based on 2D structures. Therefore, for these, conformations do not need to be generated as opposed to shape-similarity or docking approaches. This also means that 3D information will be mostly missing from the screening, although that may not impact the performance at all [67].
- It is less CPU-intensive than other methods. This means that it can be carried out in a regular computer, and with the same hardware, it will be a lot faster than other methods, especially protein-ligand docking.

Nonetheless, fingerprint-based similarity searching also has some pitfalls to be aware of:

- **Activity cliffs:** Activity cliffs are defined as pairs of compounds with very high similarity yet highly different activity, and therefore their presence can negatively impact the performance of the similarity searching. They are dependent on the dataset and the descriptors used to calculate similarity, so different approaches will show different activity cliffs in the same dataset, and finding the best solution can be tricky [68].
- **Choice of descriptors:** Similarity search performance depends greatly on the descriptors used to calculate the similarity, and in the case of fingerprints, different fingerprints can yield very different performance results [69]. The obtained results can also vary depending on the algorithm implementation.
- **Reference molecules:** For similarity searching, at least one known active molecule is needed to be used as a reference molecule. However, usually not all the parts of the reference molecules are equally relevant towards their activity, which if not taken into account may result in inactive molecules similar in irrelevant aspects to the reference molecules ranked similarly or even higher than actually active molecules which are only similar to the reference molecules in the activity-relevant aspects. A proper fingerprint choice based on the knowledge of the reference compounds may help alleviate this problem.
- **Conformation coverage:** When using 3D fingerprints, the conformations for each molecule should adequately cover its conformational space, which requires the testing and optimization of several parameters [70].

In addition, there are also many other pitfalls that are not specific to similarity searching but common for almost all virtual screening methods, as thoroughly explained by Scior et al. [70].

5. Conclusion

There are many types of fingerprints, and thus there is also interest in knowing which fingerprints perform better. There are open-source platforms to benchmark fingerprints for ligand-based virtual screening that have been tested with 14 2D fingerprints [28]. Studies have found that the overall performance of all the fingerprints was similar, though, the inter-target difference in performance was greater than the intra-target difference between fingerprints. After ranking the fingerprints by performance, these studies found that ECFP0 (with a diameter of 0 when only taking the single atom as the environment) and 166-bit MACCS were the worst when using early recognition evaluation methods. Using the same methods, circular fingerprints were ranked higher, and the topological torsions fingerprint was always highly ranked regardless of the evaluation methods.

The current trend regarding similarity searching with molecular fingerprints seems to be to combine different approaches through data fusion [71] (either by combining different fingerprints [63,72,73] or by combining fingerprints with other virtual screening methods [73,74], specially structure-based methods [75]). The advantage of this approach is that, by combining methods that capture different chemical information, the highest ranked hits will be those that are highly ranked by several approaches, making them more relevant and reducing the amount of artifacts a single approach could introduce. This could

possibly lead to the optimal search and combination of methods in data fusion, with increased virtual screening performance.

6. Acknowledgements

This manuscript was edited for English language fluency by American Journal Experts. This study was supported by grant AGL2011-25831/ALI from the Spanish Government and ACCIÓ program and XRQTC grant from ‘Generalitat de Catalunya’.

7. References

- [1] U. Rester, From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective., *Curr. Opin. Drug Discov. Devel.* 11 (2008-7) 559–68.
- [2] H. Sun, Pharmacophore-based virtual screening., *Curr. Med. Chem.* 15 (2008-1) 1018–24.
- [3] J. Kirchmair, S. Distinto, P. Markt, D. Schuster, G.M. Spitzer, K.R. Liedl, et al., How to optimize shape-based virtual screening: choosing the right query and including chemical information., *J. Chem. Inf. Model.* 49 (2009-3) 678–92. doi:10.1021/ci8004226.
- [4] J.L. Melville, E.K. Burke, J.D. Hirst, Machine learning in virtual screening., *Comb. Chem. High Throughput Screen.* 12 (2009-5) 332–43.
- [5] M. Dunkel, U. Schmidt, S. Struck, L. Berger, B. Gruening, J. Hossbach, et al., SuperScent--a database of flavors and scents., *Nucleic Acids Res.* 37 (2009-1) D291–4. doi:10.1093/nar/gkn695.
- [6] G.M. Maggiora, V. Shanmugasundaram, Molecular similarity measures., *Methods Mol. Biol.* 672 (2011-1) 39–100. doi:10.1007/978-1-60761-839-3_2.
- [7] G.M. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, Molecular Similarity in Medicinal Chemistry., *J. Med. Chem.* (2013-10). doi:10.1021/jm401411z.
- [8] H. Eckert, J. Bajorath, Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches., *Drug Discov. Today.* 12 (2007-3) 225–33. doi:10.1016/j.drudis.2007.01.011.

- [9] N. Salim, J. Holliday, P. Willett, Combination of fingerprint-based similarity coefficients using data fusion., *J. Chem. Inf. Comput. Sci.* 43 (2002-1) 435–42. doi:10.1021/ci025596j.
- [10] Accelrys, MACCS Structural Keys, (n.d.).
- [11] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL Keys for Use in Drug Discovery, *J. Chem. Inf. Model.* 42 (2002-11) 1273–1280. doi:10.1021/ci010132r.
- [12] E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant, Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities, *Annu. Rep. Comput. Chem.* 4 (2008) 217–241. doi:10.1016/S1574-1400(08)00012-1.
- [13] A. Dalke, ChemFP, (<http://chemfp.com>, accessed on 06/20/2014).
- [14] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics., *J. Chem. Inf. Comput. Sci.* 43 (2003-1) 493–500. doi:10.1021/ci025584y.
- [15] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, E.L. Willighagen, Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics., *Curr. Pharm. Des.* 12 (2006-1) 2111–20.
- [16] J.M. Barnard, G.M. Downs, Chemical Fragment Generation and Clustering Software, *J. Chem. Inf. Model.* 37 (1997-1) 141–142. doi:10.1021/ci960090k.
- [17] A. Tovar, H. Eckert, J. Bajorath, Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of

- increasing structural diversity., *ChemMedChem*. 2 (2007) 208–17.
doi:10.1002/cmdc.200600225.
- [18] R.P. Sheridan, M.D. Miller, D.J. Underwood, S.K. Kearsley, Chemical Similarity Using Geometric Atom Pair Descriptors, *J. Chem. Inf. Model.* 36 (1996-1) 128–136. doi:10.1021/ci950275b.
- [19] Chemical Computing Group Inc., Molecular Operating Environment (MOE), (2013).
- [20] I. Daylight Chemical Information Systems, Daylight, (<http://www.daylight.com/>, accessed on 06/01/2014).
- [21] OpenEye Scientific Software, OEChem, (2013).
- [22] A. Bender, H.Y. Mussa, R.C. Glen, S. Reiling, Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier., *J. Chem. Inf. Comput. Sci.* 44 (2003-1) 170–8. doi:10.1021/ci034207y.
- [23] A. Bender, H.Y. Mussa, R.C. Glen, S. Reiling, Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance., *J. Chem. Inf. Comput. Sci.* 44 (2004-1) 1708–18. doi:10.1021/ci0498719.
- [24] N.M. O’Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox., *J. Cheminform.* 3 (2011-1) 33. doi:10.1186/1758-2946-3-33.
- [25] G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner, A. Zell, jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints., *J. Cheminform.* 3 (2011-1) 3. doi:10.1186/1758-2946-3-3.

- [26] H.L. Morgan, The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service., *J. Chem. Doc.* 5 (1965-5) 107–113. doi:10.1021/c160017a018.
- [27] D. Rogers, M. Hahn, Extended-connectivity fingerprints., *J. Chem. Inf. Model.* 50 (2010-5) 742–54. doi:10.1021/ci100050t.
- [28] S. Riniker, G.A. Landrum, Open-source platform to benchmark fingerprints for ligand-based virtual screening., *J. Cheminform.* 5 (2013-1) 26. doi:10.1186/1758-2946-5-26.
- [29] Accelrys, Accelrys - Scientific Enterprise Software for Chemical Research, Material Science R&D, (<http://accelrys.com/>, accessed on 06/18/2014).
- [30] ChemAxon – cheminformatics platforms and desktop applications, (<https://www.chemaxon.com/>, accessed on 06/18/2014).
- [31] G. Landrum, RDKit: Open-source cheminformatics, (<http://www.rdkit.org>, accessed on).
- [32] Tripos :: A CertaraTM Company, (<http://tripos.com/index.php>, accessed on 06/18/2014).
- [33] L. Xue, J.W. Godden, F.L. Stahura, J. Bajorath, Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme., *J. Chem. Inf. Comput. Sci.* 43 (2003-1) 1151–7. doi:10.1021/ci030285+.
- [34] M.J. McGregor, S.M. Muskal, Pharmacophore fingerprinting. 2. Application to primary library design., *J. Chem. Inf. Comput. Sci.* 40 (1999) 117–25.
- [35] D. Vidal, M. Thormann, M. Pons, LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular

- similarities., *J. Chem. Inf. Model.* 45 (2005-1) 386–93.
doi:10.1021/ci0496797.
- [36] J. Schwartz, M. Awale, J.-L. Reymond, SMIfp (SMILES fingerprint) chemical space for virtual screening and visualization of large databases of organic molecules., *J. Chem. Inf. Model.* 53 (2013-8) 1979–89.
doi:10.1021/ci400206h.
- [37] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Model.* 29 (1989-5) 97–101. doi:10.1021/ci00062a008.
- [38] Z. Deng, C. Chuaqui, J. Singh, Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions., *J. Med. Chem.* 47 (2004-1) 337–44.
doi:10.1021/jm030331x.
- [39] M. Sastry, J.F. Lowrie, S.L. Dixon, W. Sherman, Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments., *J. Chem. Inf. Model.* 50 (2010-5) 771–84.
doi:10.1021/ci100062n.
- [40] N.M. O’Boyle, G.R. Hutchison, Cinfony – combining Open Source cheminformatics toolkits behind a common interface, *Chem. Cent. J.* 2 (2008) 24.
- [41] Open Babel, (<http://openbabel.org>, accessed on 06/20/2014).
- [42] R.E. Carhart, D.H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure-activity studies: definition and applications, *J. Chem. Inf. Model.* 25 (1985-5) 64–73. doi:10.1021/ci00046a002.
- [43] R. Nilakantan, N. Bauman, J.S. Dixon, R. Venkataraghavan, Topological torsion: a new molecular descriptor for SAR applications. Comparison

- with other descriptors, *J. Chem. Inf. Model.* 27 (1987-5) 82–85.
doi:10.1021/ci00054a008.
- [44] The Chemistry Development Kit, (<http://sourceforge.net/projects/cdk/>, accessed on 06/20/2014).
- [45] L.H. Hall, L.B. Kier, Electrotological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information, *J. Chem. Inf. Model.* 35 (1995-11) 1039–1045.
doi:10.1021/ci00028a014.
- [46] J. Klekota, F.P. Roth, Chemical substructures that enrich for biological activity., *Bioinformatics.* 24 (2008-12) 2518–25.
doi:10.1093/bioinformatics/btn479.
- [47] Indigo – GGA Software Services., (<http://ggasoftware.com/opensource/indigo>, accessed on).
- [48] cinfony - A common API for several cheminformatics toolkits - Google Project Hosting, (<https://code.google.com/p/cinfony/>, accessed on).
- [49] J. Kiener, Molecule database framework: a framework for creating database applications with chemical structure search capability., *J. Cheminform.* 5 (2013-1) 48. doi:10.1186/1758-2946-5-48.
- [50] Canvas- Product Features, (<http://www.schrodinger.com/Canvas/>, accessed on 06/20/2014).
- [51] jCompoundMapper, (<http://jcompoundmapper.sourceforge.net/>, accessed on 06/20/2014).
- [52] A. Cereto-Massagué, L. Guasch, C. Valls, M. Mulero, G. Pujadas, S. Garcia-Vallve, DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets, *Bioinformatics.* (2012-4) 2–3.
doi:10.1093/bioinformatics/bts249.

- [53] DecoyFinder, (<http://urvnutrigenomica-ctns.github.io/DecoyFinder/>, accessed on 06/20/2014).
- [54] FLAP - (Fingerprints for Ligands and Proteins), (http://www.moldiscovery.com/soft_flap.php, accessed on 06/18/2014).
- [55] M. Sud, MayaChemTools: Home, (<http://www.mayachemtools.org/>, accessed on 07/19/2014).
- [56] The PubChem Project, (<http://pubchem.ncbi.nlm.nih.gov/>, accessed on 06/20/2014).
- [57] A.J. Williams, Public chemical compound databases., *Curr. Opin. Drug Discov. Devel.* 11 (2008-5) 393–404.
- [58] H.E. Pence, A. Williams, ChemSpider: An Online Chemical Information Resource, *J. Chem. Educ.* 87 (2010-11) 1123–1124.
doi:10.1021/ed100697w.
- [59] ChemSpider | Search and share chemistry, (<http://www.chemspider.com/>, accessed on 06/20/2014).
- [60] J.J. Irwin, B.K. Shoichet, ZINC--a free database of commercially available compounds for virtual screening., *J. Chem. Inf. Model.* 45 (n.d.) 177–82.
doi:10.1021/ci049714+.
- [61] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, R.G. Coleman, ZINC: a free tool to discover chemistry for biology., *J. Chem. Inf. Model.* 52 (2012-7) 1757–68. doi:10.1021/ci3001277.
- [62] Welcome to ZINC Is Not Commercial - A database of commercially-available compounds, (<https://zinc.docking.org/>, accessed on 06/20/2014).

- [63] M. Awale, J.-L. Reymond, A multi-fingerprint browser for the ZINC database., *Nucleic Acids Res.* (2014-4) gku379–. doi:10.1093/nar/gku379.
- [64] ZINC Browser, (<http://dcb-reymond23.unibe.ch:8080/MCSS/>, accessed on 06/20/2014).
- [65] G. Tresadern, D. Bemporad, T. Howe, A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor., *J. Mol. Graph. Model.* 27 (2009-1) 860–70. doi:10.1016/j.jmgm.2009.01.003.
- [66] G.B. McGaughey, R.P. Sheridan, C.I. Bayly, J.C. Culberson, C. Kreatsoulas, S. Lindsley, et al., Comparison of topological, shape, and docking methods in virtual screening., *J. Chem. Inf. Model.* 47 (2007-1) 1504–19. doi:10.1021/ci700052x.
- [67] V. Venkatraman, V.I. Pérez-Nueno, L. Mavridis, D.W. Ritchie, Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods., *J. Chem. Inf. Model.* 50 (2010-12) 2079–93. doi:10.1021/ci100263p.
- [68] M. Cruz-Monteaagudo, J.L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M.N.D.S. Cordeiro, F. Borges, Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?, *Drug Discov. Today.* (2014-2). doi:10.1016/j.drudis.2014.02.003.
- [69] A. Bender, J.L. Jenkins, J. Scheiber, S.C.K. Sukuru, M. Glick, J.W. Davies, How similar are similarity searching methods? A principal component analysis of molecular descriptor space., *J. Chem. Inf. Model.* 49 (2009-1) 108–19. doi:10.1021/ci800249s.
- [70] T. Scior, A. Bender, G. Tresadern, J.L. Medina-Franco, K. Martínez-Mayorga, T. Langer, et al., Recognizing pitfalls in virtual screening: a

- critical review., *J. Chem. Inf. Model.* 52 (2012-4) 867–81.
doi:10.1021/ci200528d.
- [71] P. Willett, Fusing similarity rankings in ligand-based virtual screening., *Comput. Struct. Biotechnol. J.* 5 (2013-1) e201302002.
doi:10.5936/csbj.201302002.
- [72] A. Ahmed, F. Saeed, N. Salim, A. Abdo, Condorcet and borda count fusion method for ligand-based virtual screening., *J. Cheminform.* 6 (2014-1) 19. doi:10.1186/1758-2946-6-19.
- [73] P. Willett, Combination of similarity rankings using data fusion., *J. Chem. Inf. Model.* 53 (2013-1) 1–10. doi:10.1021/ci300547g.
- [74] G.M. Sastry, V.S.S. Inakollu, W. Sherman, Boosting virtual screening enrichments with data fusion: coalescing hits from two-dimensional fingerprints, shape, and docking., *J. Chem. Inf. Model.* 53 (2013-7) 1531–42. doi:10.1021/ci300463g.
- [75] F. Broccatelli, N. Brown, Best of Both Worlds: On the Complementarity of Ligand-Based and Structure-Based Virtual Screening, *J. Chem. Inf. Model.* (2014-5) 140530101617007. doi:10.1021/ci5001604.

8. Appendix

Table 1

Some similarity coefficients and distances used with fingerprints.

| Measure | Expression | Range |
|---------------------------------------|------------------------|----------|
| Tanimoto/Jaccard coefficient | $\frac{c}{a+b-c}$ | 0 to 1 |
| Euclidean distance | $\sqrt{a+b-2c}$ | 0 to N |
| City-Block/Manhattan/Hamming distance | $a+b-2c$ | 0 to N |
| Dice coefficient | $\frac{2c}{a+b}$ | 0 to 1 |
| Cosine similarity | $\frac{c}{\sqrt{ab}}$ | 0 to 1 |
| Russell-Rao coefficient | $\frac{c}{m}$ | 0 to 1 |
| Forbes coefficient | $\frac{cm}{ab}$ | 0 to 1 |
| Soergel distance | $\frac{a+b-2c}{a+b-c}$ | 0 to 1 |

Where, given the fingerprints of two compounds, A and B , m equals the total amount of bits present in the fingerprints, a equals the amount of bit set to 1 in A , b equals the amount of bits set to 1 in B and c equals the amount of bits set to 1 in both A and B .

Figures:

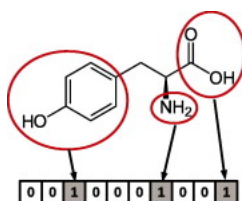


Figure 1: A representation of an hypothetical 10-bit substructure fingerprint, with three bits set because the substructures they represent are present in the molecule (circled).

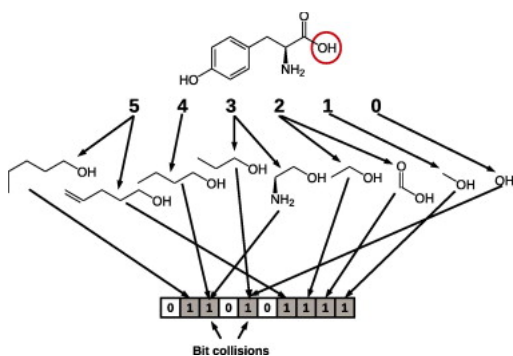


Figure 2: A representation of an hypothetical 10-bit topological fingerprint, in this case a linear path-based fingerprint with fragments up to a length of 5. All fragments which can be found from the starting atom (circled) are shown indicating the fragment length and the corresponding bit in the fingerprint. There are two bit collisions, which are bits that are set by more than one fragment, which are likely in fingerprints with a reduced amount of bits. Only fragments and bits for a single starting atom are shown, for the full fingerprint this process would be carried for every atom in the molecule. Circular fingerprints use a similar approach, but building fragments within a radius of the starting atom instead of linear fragments.

DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets

Adrià Cereto-Massagué¹, Laura Guasch¹, Cristina Valls¹, Miquel Mulero¹, Gerard Pujadas^{1,2} and Santiago Garcia-Vallvé^{1,2*}

¹Grup de Recerca en Nutrigenòmica, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus de Sescelades, C/ Marcel·lí Domingo s/n, 43007 Tarragona, Catalonia, Spain

Abstract

Summary: Decoys are molecules that are presumed to be inactive against a target (*i.e.*, will not likely bind to the target) and are used to validate the performance of molecular docking or a virtual screening workflow. The DUD database (<http://dud.docking.org/>) provides a free directory of decoys for use in virtual screening, though it only contains a limited set of decoys for 40 targets.

To overcome this limitation, we have developed an application called DecoyFinder that selects, for a given collection of active ligands of a target, a set of decoys from a database of compounds. Decoys are selected if they are similar to active ligands according to five physical descriptors (molecular weight, number of rotational bonds, total hydrogen bond donors, total hydrogen bond acceptors and the octanol-water partition coefficient) without being chemically similar to any of the active ligands used as an input (according to the Tanimoto coefficient between MACCS fingerprints). To the best of our knowledge, DecoyFinder is the first application designed to build target-specific decoy sets.

Availability: DecoyFinder is freely available at <http://URVnutrigenomica-CTNS.github.com/DecoyFinder>

Contact: santi.garcia-vallve@urv.cat

Supplementary information: A complete description of the software is included on the application home page. Validation of DecoyFinder on 10 DUD targets is provided as supplementary table 1

Introduction

Ligand enrichment is a key metric for assessing the performance of molecular docking or virtual screening workflows. It involves measuring the ability of a method or procedure to discriminate between active and inactive compounds. However, sufficient amounts of inactive compounds are generally not available for such testing; thus, decoys (i.e., molecules that are presumed to be inactive against the examined target) are commonly used for this purpose (Kirchmair et al., 2008). To avoid bias and to ensure that the enrichment is not simply due to physical differences between active and decoy compounds, decoys should exhibit physical properties (e.g., molecular weight and calculated LogP values) that are similar to active compounds, while still being chemically distinct from them (Huang et al., 2006). The largest publicly accessible database of decoys is the Directory of Useful Decoys (DUD) (Huang et al., 2006; Irwin, 2008), which is available at <http://dud.docking.org/>. The DUD contains known active and decoy compounds for 40 target proteins and is currently the gold standard for benchmarking virtual screening and molecular docking algorithms. However, the DUD only contains decoys for a small set of protein targets and has several limitations, such as the possibility of identifying a larger decoy set and the risk of overfitting (i.e., inadvertently tuning algorithms and score functions to perform well on a single benchmark) (Irwing, 2008; Wallach and Lilien, 2011). To overcome these limitations, we have created an application called DecoyFinder that selects, for a collection of active ligands of a protein target, a set of decoys from a database of compounds. To the best of our knowledge,

DecoyFinder is the first application that is designed to build target-specific decoy sets.

Program overview

Input Files: The input files that are used by DecoyFinder contain a set of active molecules (called queries) for a particular target and additional files containing a set of molecules (called potential decoys) from which decoys will be selected. These files can be in sdf, mol or any other format that is recognized by OpenBabel (<http://openbabel.org>) (O'Boyle et al., 2011), including compressed files. For the potential decoy set, the program is able to directly use subsets of the ZINC database (Irwin and Shoichet, 2005) and provides the option, if enabled, to store these subsets as cache files and use them several times. This database, available at <http://zinc.docking.org>, is free and contains over 14 million commercially available compounds for virtual screening. To avoid bias when reading the potential decoy files and to enable the acquisition of different decoy sets when DecoyFinder is re-run, potential decoy files are read in a different random order each time. In addition, it is possible to use a third file input option to submit files containing a set of known decoy molecules or decoys that have been previously selected (called known decoys) using the "add new decoys" function. These known decoy compounds will not be re-evaluated to determine whether they are decoys, but will be considered when searching for new decoys and will be included in the resulting decoy set.

Algorithm for decoy selection: The algorithm for decoy selection implemented in DecoyFinder is similar to that used to construct the DUD database (Huang et al., 2006; Irwin, 2008) and other benchmarks (Wallach and Lilien, 2011). MACCS fingerprints (Durant, 2002) and five physical descriptors are calculated

for each active and potential decoy molecule using the OpenBabel toolbox (O'Boyle et al., 2011). The Tanimoto coefficients between the MACCS fingerprints of each potential decoy and active molecule and between the potential decoys are then calculated. For each active molecule included in the query, DecoyFinder selects a set of decoys (36 when the default program options are used) from either the ZINC database or any set of molecules that is used as an input. Molecules are considered to be decoys if the following conditions are met:

They are similar to the active molecule according to five physical descriptors: molecular weight, the number of rotational bonds, total hydrogen bond donors (HBD), total hydrogen bond acceptors (HBA) and the octanol-water partition coefficient (LogP). Thus, the decoy compounds exhibit physical properties that are similar to active compounds, which prevents bias and ensures that the enrichment is not simply due to physical differences between the active and decoy compounds. Using the default program options, the physical descriptors of a decoy are considered to be similar to those of an active ligand if the following conditions are met: (i) the molecular weight is within 25 Da of the active ligand; (ii) they contain the same number ± 1 of rotational bonds and HBDS, and the same number ± 2 of HBAs; and (iii) the LogP value is within 1.0 of the active ligand. These constraint values can be relaxed in cases where a full decoy set cannot be generated or would take too much time to complete.

The Tanimoto coefficients between a potential decoy and each of the active molecules are not greater than a defined threshold (with the default set to 0.75). Thus, decoys are chemically different from any of the active molecules of the query.

The Tanimoto coefficients between a potential decoy and previously selected decoys are not greater than a defined threshold (with the default set to 0.9). This

reduces the incidence of analogous structures between decoys and the bias of analogue or trivial enrichment when decoys are used in a virtual screening workflow validation (Irwin, 2008).

As a validation, an analysis of the performance of the decoys obtained with DecoyFinder when using GlideSP to score actives and decoys for 10 DUD targets can be found as supplementary table 1.

Output: The output of DecoyFinder is an sdf file containing the decoy molecules for a specific target and a CSV file that contains information regarding the sdf file and the decoy search options. When a full decoy set cannot be generated, the program displays a warning message and redirects the output to the input screen of the “add new decoys” option. Thus, the user can attempt to complete the decoy set by either using a different library of potential decoy compounds or relaxing the constraints used.

Implementation and system requirements

DecoyFinder has been developed as a python GUI application. It has the following dependences:

- Version 4.6 or higher of Nokia’s Qt framework (<http://qt.nokia.com>). DecoyFinder uses this framework for its graphical user interface.
- OpenBabel (<http://openbabel.org>) version 2.3.0 or higher with python bindings (O’Boyle *et al.*, 2008; O’Boyle *et al.*, 2011). Prior versions contained a bug that prevented DecoyFinder from working. OpenBabel is a powerful cheminformatics toolkit that we use to parse molecule files and calculate molecular properties.
- Python version 2.6 or higher (but lower than version 3.0).
- Python Qt bindings: either PySide 1.0 or higher, or PyQt4.

A version of DecoyFinder for Ubuntu 10.10 (and newer versions), another one for Fedora 16 and a Windows version that includes all the dependencies, as well as the source code and several tools (e.g., a Wiki, documentation and a bug tracking system), are available at <http://URVnutrigenomica-CTNS.github.com/DecoyFinder>. Contextual help is provided to guide users through the set up of a DecoyFinder search run.

Acknowledgements

This manuscript has been edited by American Journal Experts.

Funding: This work was supported by the "Ministerio de Educación y Ciencia" of the Spanish Government [AGL2008-01310 and AGL2011-25831] and the ACCIÓ program from the "Generalitat de Catalunya" [TECCT11-1-0012]. We acknowledge support from the Generalitat de Catalunya through grant XRQTC.

References

- Huang,N. *et al.* (2006) Benchmarking sets for molecular docking. *J. Med. Chem.*, **49**, 6789-6801.
- Durant,J.L. *et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, **42**, 1273-1280.
- Irwin,J.J. (2008) Community benchmarks for virtual screening. *J. Comput. Aided Mol. Des.*, **22**, 193-199.
- Irwin,J.J. and Shoichet,B.K. (2005) ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, **45**, 177-182.
- Kirchmair,J. *et al.* (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy

selection--what can we learn from earlier mistakes? *J. Comput. Aided Mol. Des.*, **22**, 213-228.

O'Boyle,N.M. *et al.* (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 5.

O'Boyle,N.M. *et al.* (2011) Open Babel: An open chemical toolbox. *J. Cheminf.*, **3**, 33.

Wallach,I. and Lilien,R. (2011) Virtual decoy sets for molecular docking benchmarks. *J. Chem. Inf. Model.*, **51**, 196-202.

Molecular weight-based decoys: a simple decoy set finding alternative for fingerprint similarity approaches

Adrià Cereto-Massagué,¹ María José Ojeda,¹ Aleix Gimeno,¹ Sarah Tomas-Hernández,¹ Raúl Beltrán-Debón,¹ Josep M. Mateo-Sanz,² Cristina Valls,¹ Miquel Mulero,¹ Gerard Pujadas,^{1,3} Santiago Garcia-Vallve^{1,3}*

¹. Research Group in Cheminformatics & Nutrition. Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus de Sescelades, 43007 Tarragona, Catalonia, Spain

². Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av. Països Catalans 26, Campus de Sescelades, 43007 Tarragona, Catalonia, Spain.

³. Centre Tecnològic de Nutrició i Salut (CTNS), TECNIO, CEICS, Avinguda Universitat 1, 43204 Reus, Catalonia, Spain

* Corresponding author: santi.garcia-vallve@urv.cat

Keywords: Virtual Screening, Decoys, Decoy Sets, Validation, Molecular Fingerprints, Ligand-Based Virtual Screening

Abbreviations

VS, virtual screening; MW, Molecular weight; EF1, Enrichment factor at 1%; EF10, Enrichment factor at 10%; FP, Fingerprint; AUC, area under curve; AUROC, area under ROC curve; DEKOIS, demanding evaluation kits for

objective *in silico* screening; DUD, directory of useful decoys; DUD-E, DUD-enhanced; ROC, receiver operating characteristic; BEDROC, Boltzmann-Enhanced Discrimination of ROC; ECFP4, extended-connectivity fingerprints of maximum diameter 4.

Abstract

Decoys are molecules that are presumed to be inactive against a target (i.e. are not likely to bind to the target) and are used to validate the performance of virtual screening workflows. Current methodologies for building decoy sets for one specific target ensure that these are not chemically similar to the target's known active molecules, but are physically similar. Moreover, most decoy sets are currently built with docking in mind, and thus use some molecular fingerprints to control the similarity between actives and decoys. This could introduce an important bias if using those decoy sets in validating a virtual screening workflow which uses molecular fingerprints. In this paper, the targets in the DUD-E and DEKOIS 2.0 databases were used to analyze whether randomly selecting molecules for a decoy set with a molecular weight similar to that of the corresponding active molecules can provide useful decoy sets for molecular fingerprint virtual screening, and how do these compare to the decoy sets in DUD-E and DEKOIS 2.0, which used fingerprints for their generation. Also, using the molecular weight relative to the target's active molecules as the only criterion for selecting the molecules for the corresponding decoy set is

very fast and, therefore, makes it possible to obtain decoys for almost any drug target with known active molecules. Consequently, we have adapted our DecoyFinder software so that this criterion can be used as the default when obtaining decoy data sets.

Introduction

Virtual screening (VS) is a computational method used to search large libraries of molecules and identify those structures that are most likely to bind a specific target and modulate its bioactivity in a pre-defined way.¹ Validation is a key aspect of a VS workflow, as any new protocol needs to prove that it can actually discern between active compounds and inactive molecules.² The best validation method, of course, is to subject every VS hit to *in vitro* testing and then compare the method's predictions and the actual laboratory results. However, this is impractical and extremely expensive when there are hundreds or even thousands of compounds to assay, so *in silico* validation methods are used instead.² The minimum requirement of *in silico* validation methods is that one or more active molecules be known, and that some inactive molecules be known, suspected or assumed. If the VS method can successfully identify the already known active molecules from the rest, it is regarded as validated. However, this approach has some problems: on one hand, if the inactive (or assumed inactive) molecules are too different from the active molecules, telling them apart would be trivial and therefore not indicative of the potential of the method for actual VS in molecular databases (where active and inactive molecules can be quite similar).^{3,4} One way to minimize this problem is to use 'decoy' molecules, which are physically similar to active molecules, yet

chemically different, but are unlikely to be active.³⁻⁵ On the other hand, a molecule that is assumed to be inactive could unknowingly be active, and wrongly decrease the apparent performance of the method being validated because it is a false positive. One way to minimize the effect of a possible unknown active among the decoys is to use so many decoys – the most common ratio is 30-100 decoys for every active molecule – that false negatives become irrelevant.³⁻⁵ These decoy molecules are usually selected on the basis of physicochemical molecular descriptors of the known actives such as similar molecular weight (MW), number of rotatable bonds, octanol-water partition coefficient, charges, and hydrogen bond donors and acceptors. Dissimilarity to the actives, however, is guaranteed through the use of some sort of molecular fingerprint³⁻⁵. Unfortunately, this could be a source of bias when planning to use such decoys to validate a molecular fingerprint-based VS protocol.

Some databases^{3,4,6} have active and decoy sets for several targets, and there are also tools for generating custom decoy sets for custom active sets.^{5,7} Because of the constraints by which they are built, however, these decoy sets may not be representative of a real-world case of VS because molecular databases also contain inactive molecules that are chemically similar to active compounds. Therefore, we tested the performance of these decoy sets^{3,4} derived from complex constraints and processes, including similarity restrictions through molecular fingerprints, compared to same-size sets of ‘decoy’ molecules, whose only constraint is the MW (as a radically different MW between active and decoy would make discarding the decoys trivial, thus leading to artificial enrichment) and deliberately avoiding the use of fingerprints to limit the potential active-decoy similarities. The comparison was between the

performances of a VS process by using different molecular fingerprint methods and metrics.

Computational methods

Two decoy databases were taken as references and their active and original decoy sets used: DUD-E³ (101 different targets) and DEKOIS 2.0⁴ (74 different targets). For each target in DUD-E and DEKOIS 2.0 a new set of decoys was obtained with molecules from the ZINC⁸ database. The only constraint for selecting one ZINC molecule to be part of this new set of decoys is that their MW must be within one standard deviation of the MW of all the actives of the corresponding target. This method is much simpler than those used by DUD-E, DEKOIS 2.0 and others^{9,10} for generating decoy sets. The decoy sets were generated so they had the same active to decoy ratio as the original database, which was 50 decoys per active for DUD-E and 30 decoys per active for DEKOIS 2.0.

For each target from DUD-E and DEKOIS 2.0, a set of active ligands was selected from PDB models in which the ligand was co-crystallized with the target. The reliability of these ligand coordinates at the different models was validated using VHELIBS.¹¹ These ligand sets were used as references during fingerprint comparisons with actives and decoys from the corresponding target. All actives from the original active sets that had the same InChI¹² as any of the reference ligands for the same target were removed from their sets. The original decoy sets were left untouched.

In parallel, for each target several molecular fingerprints were used to calculate the Tanimoto coefficient between the reference ligands and: (1) the actives; (2) the original decoy set; and (3) our MW-based decoy set. During this comparison, the score assigned to each active or decoy was its highest Tanimoto value with any of the reference ligands. On the basis of this value, a ROC curve was calculated and the performance was measured using several metrics: AUROC (area under the ROC curve), BEDROC¹³ ($\alpha = 20$) and enrichment factors at 1% and 10% of the decoys (as used by DUD-E³). Performance of the decoy sets was tested using several different molecular fingerprints available in ChemFP¹⁴: MACCS166¹⁵ (OpenEye¹⁶, OpenBabel¹⁷ and RDKit¹⁸ implementations), OpenEye Path, Circular and Tree; RDKit's path-based, Morgan (ECFP4-like),^{18,19} Torsion, and AtomPair; OpenBabel's FP2, FP3 and FP4; and ChemFP's implementations of the Pubchem fingerprint on top of OpenBabel, RDKit and OpenEye toolkits.

The performance results of the original and the MW-based decoys were then statistically compared for each database by means of a linear regression for each fingerprint type and their differences were assessed. For each of the linear regressions performed, the correlation coefficient, mean performances of both sets, mean difference, slope, intersect and P-value are provided in the supplementary information. BEDROC and EF10% results are shown in Tables 1-2 and 3-4, respectively.

Results and discussion

Original DUD-E decoy sets were generated by property matching ZINC molecules to the active ligands using MW, estimated octanol-water partition coefficient, number of rotatable bonds, hydrogen bond donors and acceptors and their net charge. The representative states of each ligand were in the pH range 6-8, and the most dissimilar of these was chosen according to an ECFP4 fingerprint.³ Original DEKOIS 2.0 decoy sets were generated from ZINC database molecules taking into account their similarity to the actives in terms of MW, estimated octanol-water partition coefficient, hydrogen bond donors and acceptors, number of rotatable bonds, positive charges, negative charges and aromatic rings, all of which were calculated at pH 7.4. The presence of latent actives in the decoy sets of DEKOIS 2.0 was minimized by using a custom score based on FCFP6 fingerprints to avoid potentially bioactive structures.⁴ Therefore, when their performance is compared to a set of decoys that has been generated without incorporating any specific measure against analogue bias or artificial enrichment (apart from the MW restriction), the DEKOIS 2.0 and DUD-E decoy sets would be expected to yield a much higher VS performance because of the artificial enrichment and other effects. However, the results show that the decoys generated by only taking the MW into account have a very similar fingerprint-based VS performance to those generated by the two more complex methods, regardless of the metric (see Figure 1). Thus, the slope of a linear regression between the results obtained using the original decoys and those obtained using our approach is significantly close to 1 (see Figure 1). As can be seen in Tables from 1 to 4 for BEDROC and EF10, the mean difference is in most cases in favor of the original decoys but just by a small margin. More detailed tables for every metric used and linear regression plots for every

fingerprint type used can be found in the Supporting Information. Interestingly, performance differences were quite consistent across all the fingerprint types used, even when the algorithm of the fingerprints was very similar to those used in the original process for generating the decoy set and choosing the most dissimilar decoys (for example, Morgan and OpenEye Circular fingerprints provide similar results to ECFP4 and FCFP6 fingerprints, respectively). For comparison, the distribution of hydrogen bond donors, hydrogen bond acceptors and calculated log P across actives, original decoys and MW-based decoys were plotted for each target and can be found in the Supporting Information. As a summary, the log P overall distribution plots for all actives, original decoys and MW-based decoys for all DUD-E and DEKOIS targets are shown in Figures 2 and 3, respectively. There, it can be seen that MW-based decoys lie well within the values of both the actives and the original decoys, while deviating less from the average. This means that MW-based decoys may be even harder to tell apart from “average” actives than the original decoys based on calculated log P alone.

By looking at the performance figures, such as Figure 1, it can be seen that there are some outlier cases.

One such a case is the one of the Urokinase Plasminogen activator (uPA, from the DEKOIS 2.0 dataset) with OpenEye's circular fingerprint, where its performance is much higher for the molecular-weight based decoys than for its original decoys. By looking at the comparison between in physical properties between the actives and the decoy sets (Figures 4,5 and 6), one could conclude that this may be caused by the fact that the MW-based decoys, in this case, tend to have a higher calculated log P when compared both to the actives and the

original decoys. However, such a big difference in performance is not found when performing the analysis using any other of the fingerprints tested.

A case in the opposite side of the plot, showing a much lower performance with the MW-based decoy sets, would be that of the Cytochrome P450 2A6 (CYP2A6) with RDKit's AtomPair fingerprint (see Figure 1). Looking at Figures 7, 8 and 9, we can see that this time there is also a slight discrepancy in calculated log P, but in the opposite direction. Here, again, choosing a different fingerprint also negates this difference in performance.

Conclusion

The results of this study indicate that artificial enrichment due to trivial differences in physicochemical properties between actives and decoys does not play a very important role, as most of the decoy sets ignoring them perform just slightly worse instead than better than those avoiding artificial enrichment. One explanation for this could be a high percentage of latent actives among the decoy molecules in all targets, but this is highly unlikely because they were selected only on the basis of their MW and there were none other measures that could make the selected decoys more similar to the actives.

Overall, the results show that, for fingerprint-based VS, decoys that are selected using the MW relative to the target's active molecules as the only criterion for selecting the molecules, perform similarly to decoys that are selected taking many other criteria into account. The DUD-E and DEKOIS 2.0 are regarded as the maximal-unbiased benchmarking sets²⁰, and our results show that they are viable for their use in validation of fingerprint-based VS protocols without fearing for biases introduced by the use of similar molecular fingerprints in the VS and in the decoy-building procedure. However, it still remains very useful to be able to easily build decoy sets of any size for targets outside of those covered by the DUD-E and DEKOIS 2.0, or even to complement them. Another possible applicability of such decoy sets would be to test which molecular fingerprints perform better for a certain set of active molecules, with no fear of fingerprint bias, in order to select the best-suited algorithm for each case. Therefore, we have updated our decoy set building software, DecoyFinder,²¹ to take advantage of this and it is now able to find decoy sets with less computational burden.

In this study, we used the same decoy/active ratio as the databases we were comparing against, but increasing the amount of decoys for each active would most likely increase the VS performance in most cases.

Docking performance falls outside the scope and intent of this study, but further research could be done to test whether such MW-based decoys perform similarly to traditional decoys in docking VS workflows.

Tables

Table 1. Statistical analysis of the BEDROC results obtained with the MW-based decoys in comparison with the original DUD-E decoys

| BEDROC DUD-E | Correlation coefficient | Mean (Original decoys) | Mean (MW decoys) | Mean difference | P-value | Slope | Intersect |
|----------------------------|-------------------------|------------------------|------------------|-----------------|----------|-------|-----------|
| ChemFP-Substruct-OpenBabel | 0.952 | 0.543 | 0.473 | 0.070 | < 0.0001 | 0.973 | -0.055 |
| ChemFP-Substruct-OpenEye | 0.953 | 0.543 | 0.474 | 0.069 | < 0.0001 | 0.971 | -0.053 |
| ChemFP-Substruct-RDKit | 0.954 | 0.543 | 0.474 | 0.069 | < 0.0001 | 0.972 | -0.054 |
| OpenBabel-FP2 | 0.949 | 0.529 | 0.520 | 0.009 | < 0.0001 | 1.019 | -0.019 |
| OpenBabel-FP3 | 0.932 | 0.196 | 0.121 | 0.075 | < 0.0001 | 0.841 | -0.044 |
| OpenBabel-FP4 | 0.942 | 0.426 | 0.349 | 0.077 | < 0.0001 | 0.954 | -0.058 |
| OpenBabel-MACCS | 0.936 | 0.503 | 0.446 | 0.057 | < 0.0001 | 0.978 | -0.047 |
| OpenEye-Circular | 0.911 | 0.601 | 0.545 | 0.055 | < 0.0001 | 0.963 | -0.033 |
| OpenEye-MACCS166 | 0.929 | 0.494 | 0.424 | 0.070 | < 0.0001 | 0.975 | -0.058 |
| OpenEye-Path | 0.851 | 0.540 | 0.564 | -0.023 | < 0.0001 | 0.990 | 0.029 |
| OpenEye-Tree | 0.938 | 0.658 | 0.574 | 0.084 | < 0.0001 | 0.997 | -0.082 |
| RDKit-AtomPair | 0.940 | 0.536 | 0.585 | -0.049 | < 0.0001 | 1.019 | 0.039 |
| RDKit-Fingerprint | 0.959 | 0.244 | 0.297 | -0.053 | < 0.0001 | 0.988 | 0.056 |
| RDKit-MACCS166 | 0.932 | 0.504 | 0.446 | 0.058 | < 0.0001 | 0.981 | -0.048 |
| RDKit-Morgan | 0.920 | 0.668 | 0.560 | 0.108 | < 0.0001 | 0.963 | -0.083 |
| RDKit-Torsion | 0.948 | 0.646 | 0.591 | 0.055 | < 0.0001 | 1.001 | -0.056 |

Table 2. Statistical analysis of the BEDROC results obtained with the MW-based decoys in comparison with the original DEKOIS 2.0 decoys

| BEDROC DEKOIS | Correlation coefficient | Mean (Original decoys) | Mean (MW decoys) | Mean difference | P-value | Slope | Intersect |
|----------------------------|-------------------------|------------------------|------------------|-----------------|----------|-------|-----------|
| ChemFP-Substruct-OpenBabel | 0.924 | 0.470 | 0.452 | 0.018 | < 0.0001 | 0.905 | 0.026 |
| ChemFP-Substruct-OpenEye | 0.953 | 0.470 | 0.457 | 0.013 | < 0.0001 | 0.968 | 0.003 |
| ChemFP-Substruct-RDKit | 0.954 | 0.471 | 0.458 | 0.012 | < 0.0001 | 0.967 | 0.003 |
| OpenBabel-FP2 | 0.961 | 0.529 | 0.505 | 0.024 | < 0.0001 | 0.949 | 0.003 |
| OpenBabel-FP3 | 0.857 | 0.176 | 0.126 | 0.050 | < 0.0001 | 0.856 | -0.024 |
| OpenBabel-FP4 | 0.859 | 0.400 | 0.344 | 0.056 | < 0.0001 | 0.883 | -0.010 |
| OpenBabel-MACCS | 0.940 | 0.412 | 0.527 | -0.115 | < 0.0001 | 0.968 | 0.128 |
| OpenEye-Circular | 0.917 | 0.557 | 0.495 | 0.062 | < 0.0001 | 0.976 | -0.049 |
| OpenEye-MACCS166 | 0.949 | 0.411 | 0.396 | 0.015 | < 0.0001 | 0.975 | -0.005 |
| OpenEye-Path | 0.917 | 0.570 | 0.539 | 0.031 | < 0.0001 | 0.951 | -0.004 |
| OpenEye-Tree | 0.961 | 0.578 | 0.556 | 0.022 | < 0.0001 | 1.006 | -0.025 |
| RDKit-AtomPair | 0.949 | 0.517 | 0.563 | -0.046 | < 0.0001 | 0.943 | 0.075 |
| RDKit-Fingerprint | 0.861 | 0.230 | 0.320 | -0.090 | < 0.0001 | 0.963 | 0.098 |
| RDKit-MACCS166 | 0.950 | 0.413 | 0.411 | 0.002 | < 0.0001 | 1.007 | -0.005 |
| RDKit-Morgan | 0.959 | 0.568 | 0.531 | 0.037 | < 0.0001 | 0.989 | -0.031 |
| RDKit-Torsion | 0.960 | 0.575 | 0.601 | -0.025 | < 0.0001 | 0.962 | 0.047 |

Table 3. Statistical analysis of the Enrichment Factor 10% results obtained with the MW-based decoys in comparison with the original DUD-E decoys

| EF10 DUD-E | Correlation coefficient | Mean (Original decoys) | Mean (MW decoys) | Mean difference | P-value | Slope | Intersect |
|----------------------------|----------------------------|------------------------------|---------------------|--------------------|----------|-------|-----------|
| ChemFP-Substruct-OpenBabel | 0.95 | 62.88 | 55.35 | 7.53 | < 0.0001 | 10.97 | -5.75 |
| ChemFP-Substruct-OpenEye | 0.95 | 63.01 | 55.52 | 7.49 | < 0.0001 | 10.96 | -5.28 |
| ChemFP-Substruct-RDKit | 0.95 | 62.83 | 55.51 | 7.32 | < 0.0001 | 10.96 | -4.57 |
| OpenBabel-FP2 | 0.96 | 62.50 | 58.54 | 3.96 | < 0.0001 | 10.98 | -2.98 |
| OpenBabel-FP3 | 0.88 | 26.55 | 17.79 | 8.76 | < 0.0001 | 10.88 | -5.53 |
| OpenBabel-FP4 | 0.92 | 54.02 | 42.71 | 11.31 | < 0.0001 | 10.92 | -7.16 |
| OpenBabel-MACCS | 0.94 | 56.51 | 51.15 | 5.36 | < 0.0001 | 10.99 | -4.87 |
| OpenEye-Circular | 0.97 | 66.86 | 60.28 | 6.58 | < 0.0001 | 11.02 | -7.71 |
| OpenEye-MACCS166 | 0.93 | 56.24 | 49.15 | 7.09 | < 0.0001 | 10.97 | -5.47 |
| OpenEye-Path | 0.95 | 69.07 | 63.21 | 5.86 | < 0.0001 | 10.99 | -5.24 |
| OpenEye-Tree | 0.95 | 72.47 | 63.58 | 8.89 | < 0.0001 | 11.04 | -11.54 |
| RDKit-AtomPair | 0.94 | 64.33 | 65.97 | -1.64 | < 0.0001 | 10.96 | 4.47 |
| RDKit-Fingerprint | 0.93 | 27.54 | 35.23 | -7.69 | < 0.0001 | 10.97 | 8.45 |
| RDKit-MACCS166 | 0.94 | 56.77 | 51.29 | 5.48 | < 0.0001 | 10.99 | -4.72 |
| RDKit-Morgan | 0.94 | 71.51 | 61.00 | 10.51 | < 0.0001 | 11.00 | -10.67 |
| RDKit-Torsion | 0.96 | 70.76 | 66.21 | 4.55 | < 0.0001 | 11.00 | -4.72 |

Table 4. Statistical analysis of the Enrichment Factor 10% results obtained with the MW-based decoys in comparison with the original DEKOIS decoys

| EF10 DEKOIS | Correlation coefficient | Mean (Original decoys) | Mean (MW decoys) | Mean difference | P-value | Slope | Intersect |
|----------------------------|----------------------------|------------------------------|---------------------|--------------------|----------|-------|-----------|
| ChemFP-Substruct-OpenBabel | 0.89 | 55.28 | 52.43 | 2.85 | < 0.0001 | 0.85 | 5.20 |
| ChemFP-Substruct-OpenEye | 0.92 | 55.35 | 53.37 | 1.98 | < 0.0001 | 0.90 | 3.41 |
| ChemFP-Substruct-RDKit | 0.92 | 55.45 | 53.11 | 2.34 | < 0.0001 | 0.90 | 3.06 |
| OpenBabel-FP2 | 0.97 | 60.55 | 56.14 | 4.42 | < 0.0001 | 0.92 | 0.13 |
| OpenBabel-FP3 | 0.82 | 24.87 | 17.76 | 7.11 | < 0.0001 | 0.77 | -1.43 |
| OpenBabel-FP4 | 0.85 | 46.35 | 38.91 | 7.44 | < 0.0001 | 0.85 | -0.39 |
| OpenBabel-MACCS | 0.93 | 46.63 | 57.76 | -11.13 | < 0.0001 | 0.89 | 16.31 |
| OpenEye-Circular | 0.90 | 61.63 | 52.01 | 9.61 | < 0.0001 | 0.96 | -7.44 |
| OpenEye-MACCS166 | 0.92 | 47.16 | 43.85 | 3.30 | < 0.0001 | 0.92 | 0.27 |
| OpenEye-Path | 0.92 | 64.22 | 59.45 | 4.77 | < 0.0001 | 0.95 | -1.28 |
| OpenEye-Tree | 0.97 | 65.54 | 60.97 | 4.57 | < 0.0001 | 1.01 | -5.52 |
| RDKit-AtomPair | 0.94 | 56.62 | 60.73 | -4.11 | < 0.0001 | 0.92 | 8.73 |
| RDKit-Fingerprint | 0.85 | 24.37 | 36.35 | -11.98 | < 0.0001 | 0.98 | 12.43 |
| RDKit-MACCS166 | 0.96 | 46.80 | 45.75 | 1.06 | < 0.0001 | 1.00 | -0.85 |
| RDKit-Morgan | 0.94 | 62.74 | 56.53 | 6.21 | < 0.0001 | 1.00 | -6.46 |
| RDKit-Torsion | 0.96 | 64.77 | 66.37 | -1.61 | < 0.0001 | 0.95 | 4.99 |

ASSOCIATED CONTENT

The following supporting information for this article is provided: detailed statistical analysis tables for the results of each metric (AUC, BEDROC, EF1 and EF10); linear regression plots for every target, fingerprint and metric combination; and two files with a list of the PDB ligands used as reference structures and the targets for which they were used.

This material will be available free of charge via the Internet at <http://pubs.acs.org>.

Author information

Corresponding author

Santiago Garcia-Vallve: santi.garcia-vallve@urv.cat

Author contributions

AC-M, GP and SG-V conceived the study and wrote the manuscript. AC-M and MJO performed the calculations. JMM-S provided the statistics. All authors analyzed the data, discussed the results and approved the final version of the manuscript.

Funding sources

This study was supported by the Government of Catalonia (ACC1Ó program, and grants XRQTC and 2014 SGR 537) and the Spanish Government (project TIN2011-27076-C03-01 "CO-PRIVACY").

Acknowledgements

We would like to thank Dr Mark Mackey, Cresset's Chief Scientific Officer, for inspiration on the subject of this article and also OpenEye Scientific Software (Santa Fe, NM) for providing free access to its software through their Academic Licensing program. The language of the manuscript was checked by the Language Service of the Universitat Rovira i Virgili (URV).

References

- (1) Rester, U. *Curr. Opin. Drug Discov. Devel.* **2008**, *11*, 559–568.
- (2) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. *J. Med. Chem.* **2005**, *48*, 962–976.
- (3) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. *J. Med. Chem.* **2012**, *55*, 6582–6594.
- (4) Bauer, M.; Ibrahim, T. *J. Chem. ...* **2013**.
- (5) Cereto-Massagué, A.; Guasch, L.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. *Bioinformatics* **2012**, *28*, 1661–1662.

- (6) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. *J. Chem. Inf. Model.* **2010**, *50*, 2079–2093.
- (7) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.
- (8) Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model.* *45*, 177–182.
- (9) Wallach, I.; Lilien, R. *J. Chem. Inf. Model.* **2011**, *51*, 196–202.
- (10) Gatica, E. A.; Cavasotto, C. N. *J. Chem. Inf. Model.* **2012**, *52*, 1–6.
- (11) Cereto-Massagué, A.; Ojeda, M. J.; Joosten, R. P.; Valls, C.; Mulero, M.; Salvado, M. J.; Arola-Arnal, A.; Arola, L.; Garcia-Vallvé, S.; Pujadas, G. *J. Cheminform.* **2013**, *5*, 36.
- (12) IUPAC - International Union of Pure and Applied Chemistry: The IUPAC International Chemical Identifier (InChI)
<http://www.iupac.org/home/publications/e-resources/inchi.html> (accessed Oct 15, 2014).
- (13) Truchon, J.-F.; Bayly, C. I. *J. Chem. Inf. Model.* *47*, 488–508.
- (14) Dalke, A. ChemFP <http://chemfp.com> (accessed Jun 20, 2014).
- (15) Accelrys. MACCS Structural Keys.
- (16) OpenEye Scientific Software. Santa Fe, NM. <http://www.eyesopen.com>.
GraphSim TK version 2.1.1, 2013.
- (17) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform.* **2011**, *3*, 33.
- (18) Landrum, G. RDKit: Open-source cheminformatics <http://www.rdkit.org>.
- (19) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107–113.

(20) Xia, J.; Tilahun, E. L.; Reid, T.-E.; Zhang, L.; Wang, X. S. *Methods* **2014**.

(21) DecoyFinder <http://urvnutrigenomica-ctns.github.io/DecoyFinder/>
(accessed Jun 20, 2014).

Figures

Figure 1 Comparison of calculated LogP value distribution across DEKOIS2 .0 actives and decoys, and molecular-weight based decoys for all targets

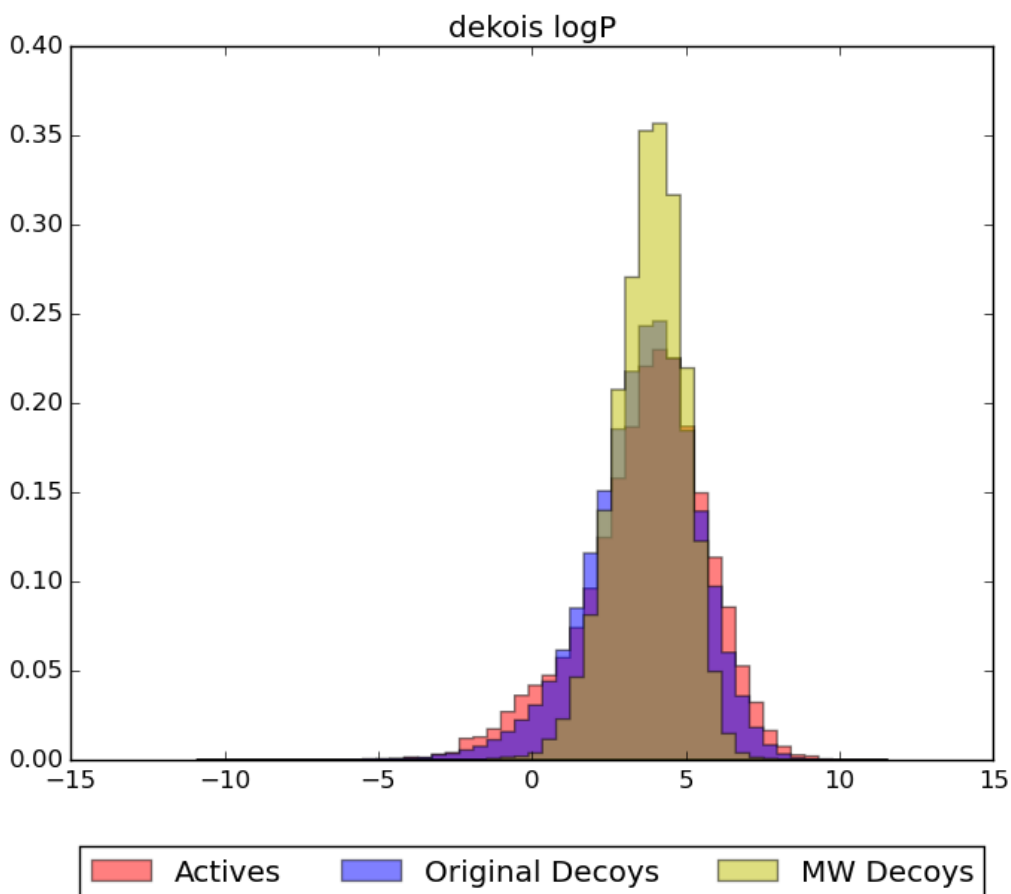


Figure 2: Hydrogen bond acceptors of actives, original decoys and MW-based decoys for target uPA

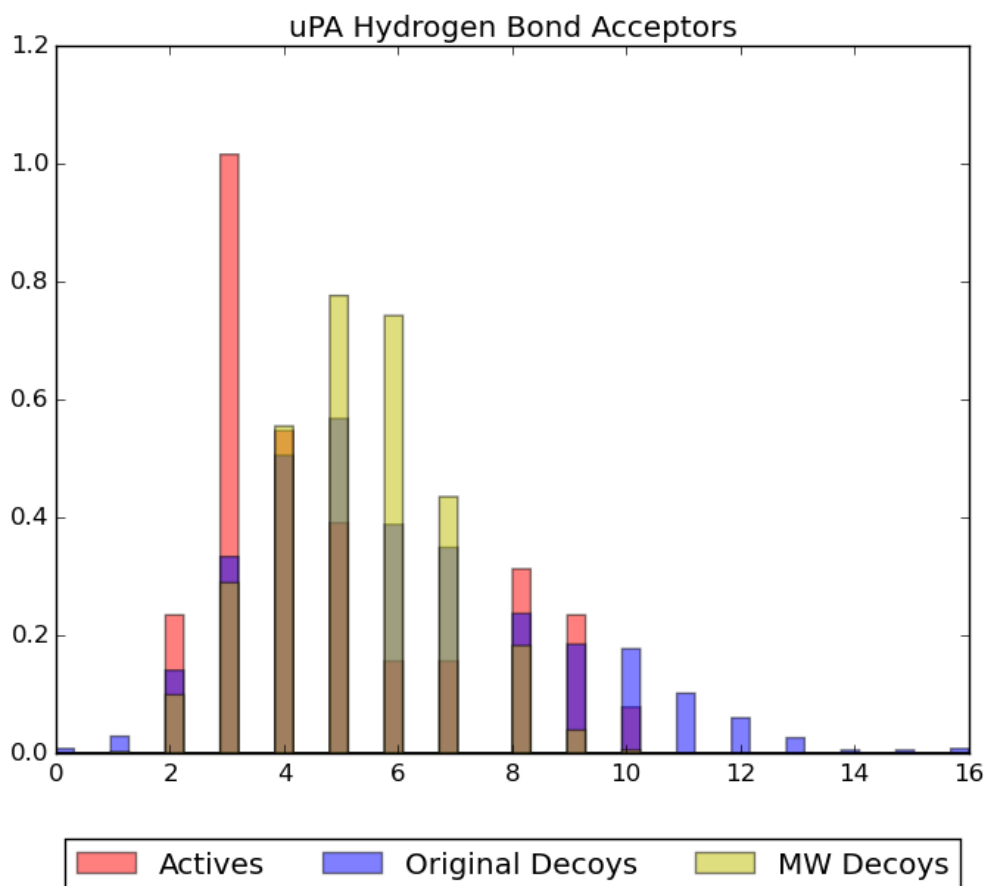


Figure 3: Hydrogen bond donors of actives, original decoys and MW-based decoys for target uPA

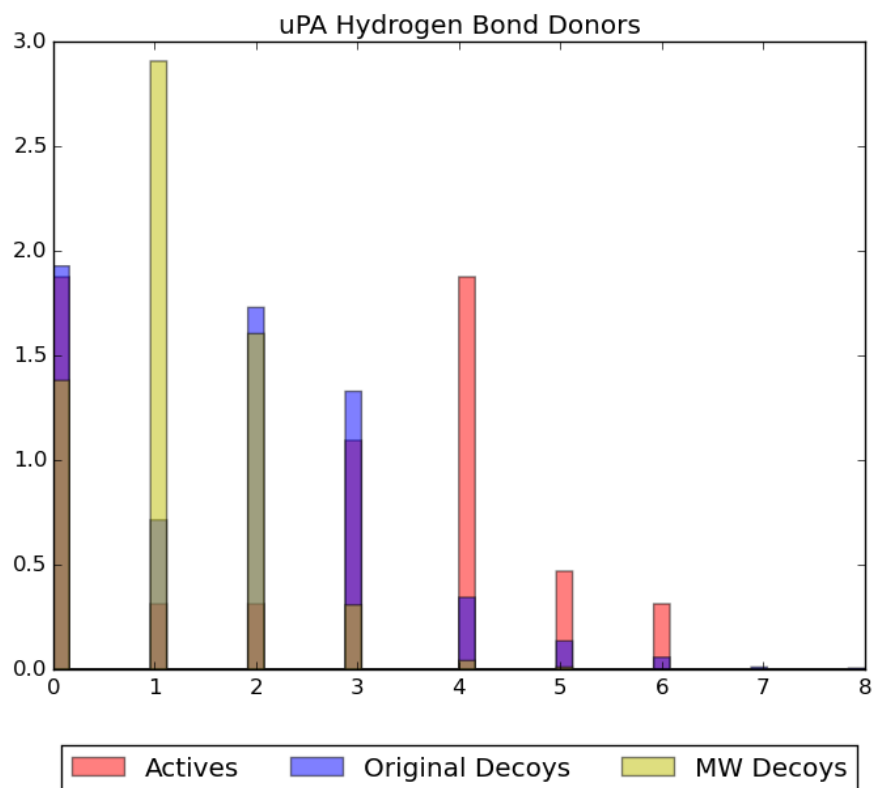


Figure 4: Calculated log P for actives, original decoys and MW-based decoys for target uPA

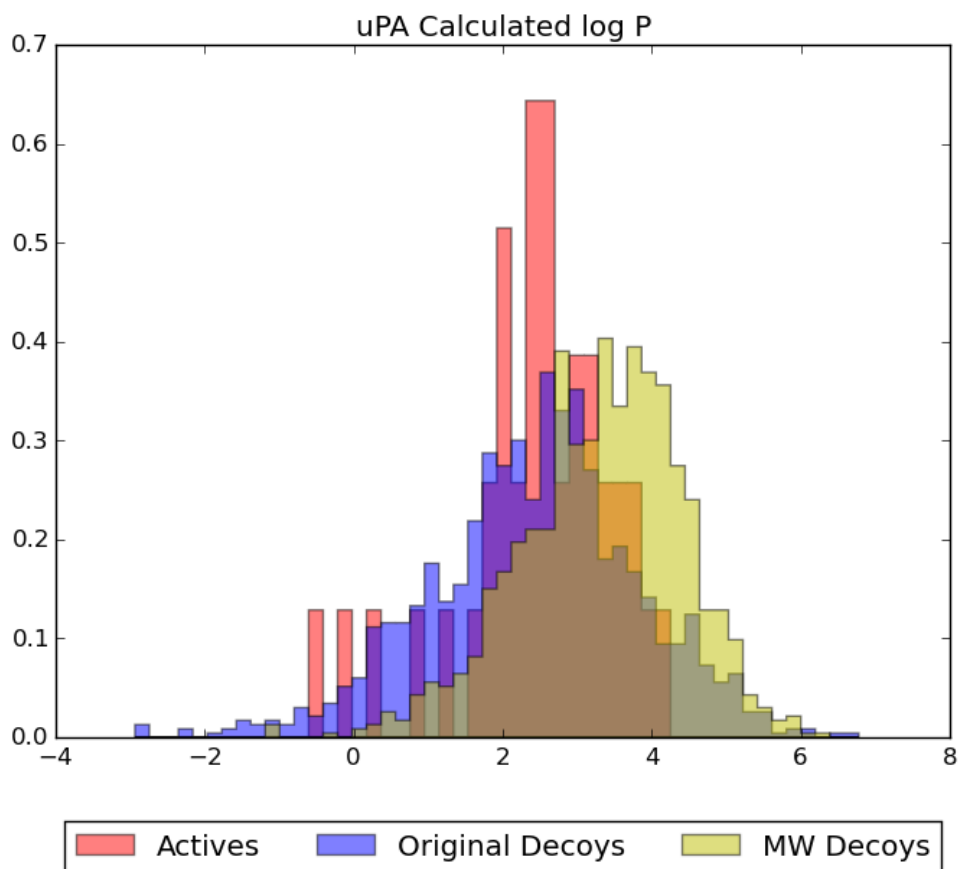


Figure 5: Hydrogen bond acceptors of actives, original decoys and MW-based decoys for target CYP2A6

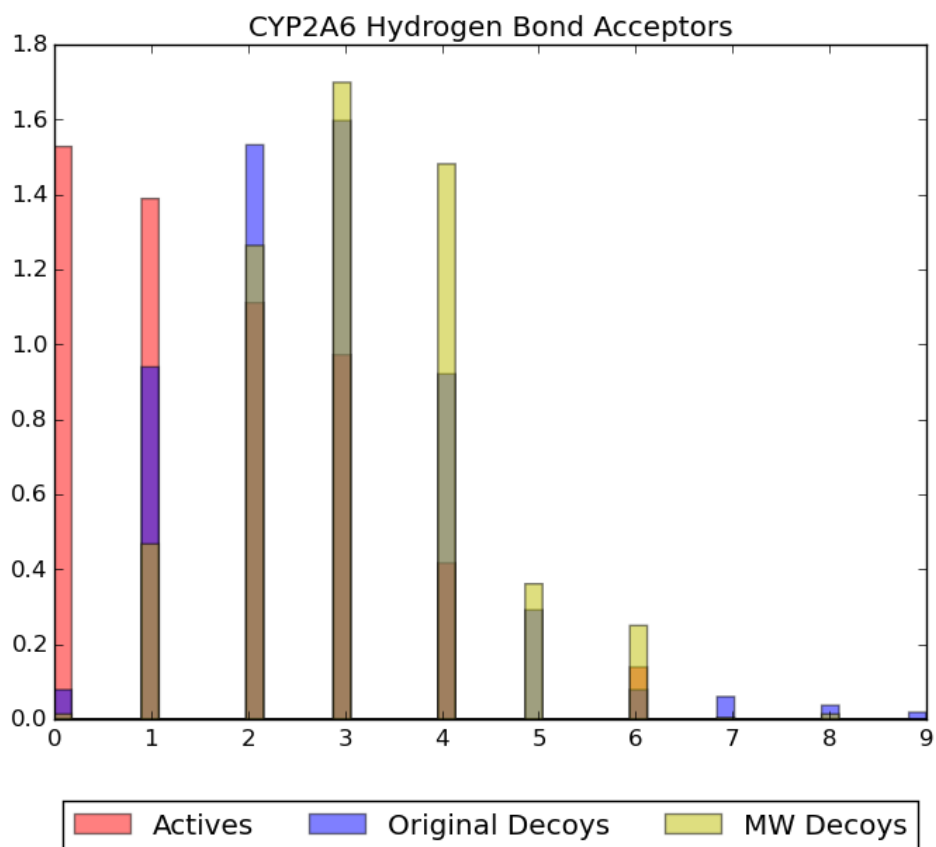


Figure 6: Hydrogen bond donors of actives, original decoys and MW-based decoys for target CYP2A6

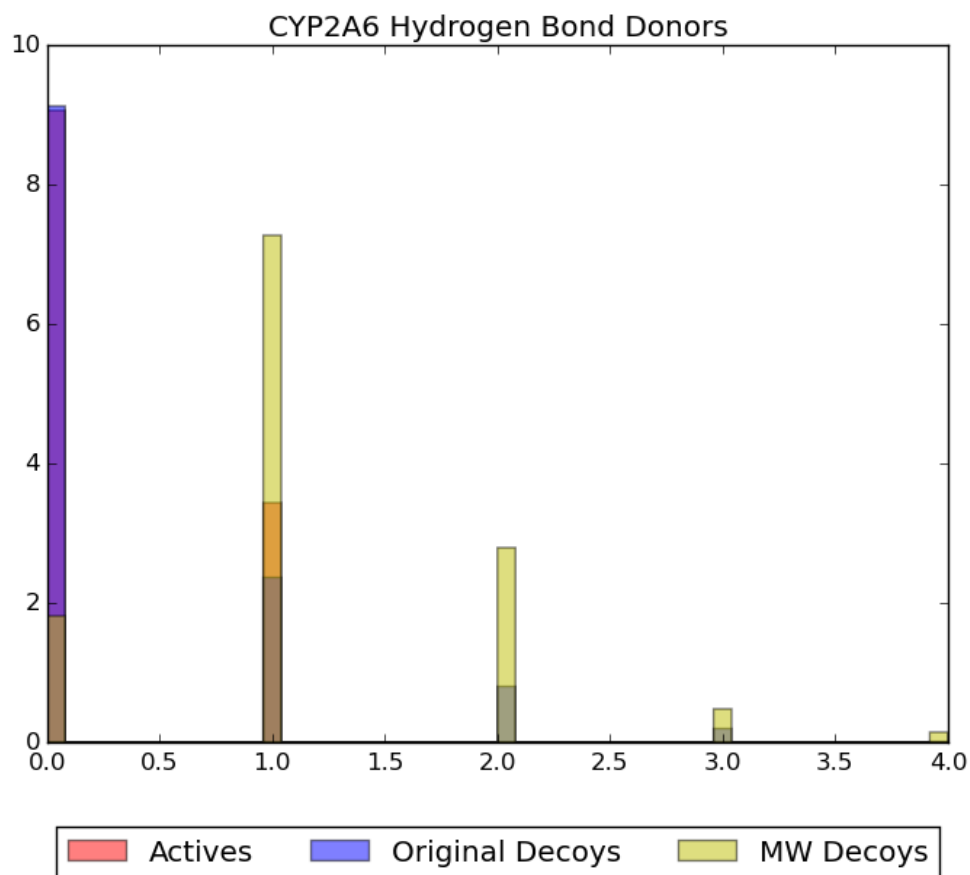
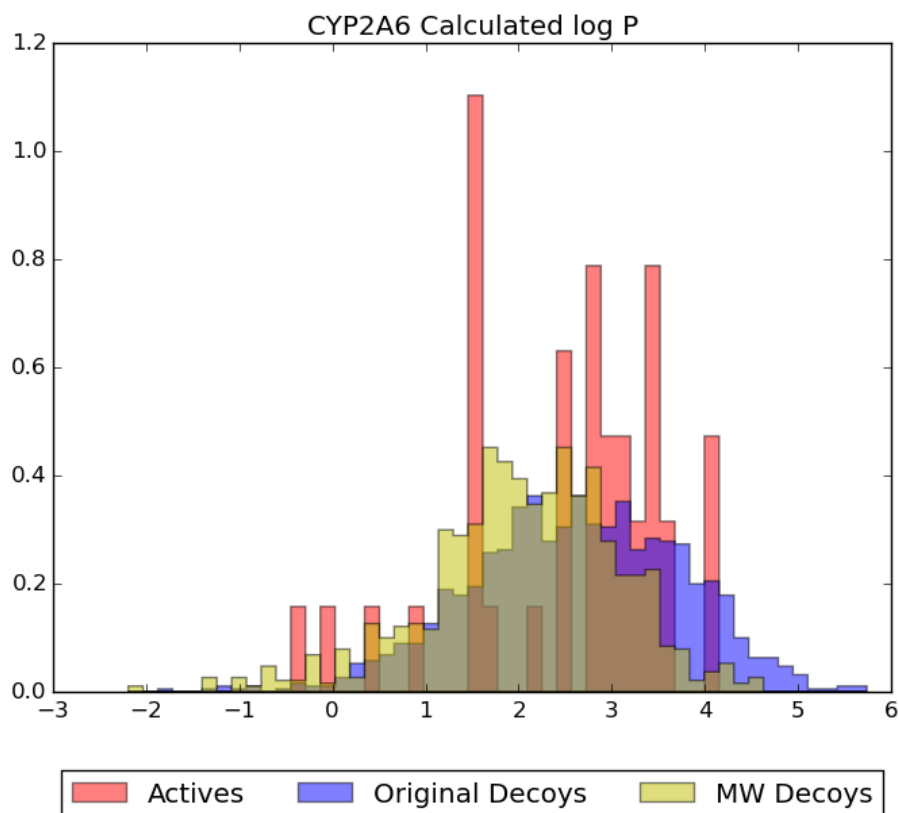


Figure 7: Calculated log P for actives, original decoys and MW-based decoys for target CYP2A6



The Good, the Bad and the Dubious. VHELIBS, a Validation Helper for Ligands and Binding Sites

Adrià Cereto-Massagué¹, María José Ojeda¹, Robbie P. Joosten², Cristina Valls¹, Miquel Mulero¹, M. Josepa Salvado¹, Anna Arola-Arnal¹, Lluís Arola^{1,3}, Anastassis Perrakis², Santiago Garcia-Vallvé^{1,3}

Corresponding author: Gerard Pujadas^{1,3,*} <gerard.pujadas@urv.cat>

¹Grup de Recerca en Nutrigenòmica, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus de Sescelades, C/ Marcellí Domingo s/n, 43007 Tarragona, Catalonia, Spain

²Department of Biochemistry, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

³Centre Tecnològic de Nutrició i Salut (CTNS), TECNIO, CEICS, Avinguda Universitat 1, 43204, Reus, Catalonia, Spain

Abstract

Background: Many Protein Data Bank (PDB) users assume high quality for the deposited structural models, but forget they are derived from the interpretation of experimental data. The accuracy of model coordinates is not homogeneous between models or throughout the same model. To avoid basing a research project on a flawed model, we present a tool for assessing the quality of ligands and binding sites in crystallographic models in the PDB.

Results: The **Validation HELper for LIgands and Binding Sites** (VHELIBS) is a software that aims to ease the validation of binding site and ligand coordinates for non-crystallographers (i.e. users with little or no crystallography knowledge). Using a convenient graphical user interface, it allows checking how ligand and binding site coordinates fit to the electron density map. VHELIBS can use models from either the PDB or the PDB_REDO databank of re-refined and re-built crystallographic models. The user can specify threshold values for a series of properties related to the fit of coordinates to electron density (Real Space R, Real Space Correlation Coefficient and average occupancy are used by default). VHELIBS will automatically classify residues and ligands as good, dubious or bad based on the specified limits. The user is able to also visually check the quality of the fit of residues and ligands to the electron density map, and reclassify them if needed.

Conclusions:

Using VHELIBS allows inexperienced users to examine the binding site and the ligand coordinates in relationship with the experimental data. This is an important step to evaluate models for their fitness for drug discovery purposes

such as structure-based pharmacophore development and protein-ligand docking experiments.

Keywords

Electron density map, binding site structure validation, ligand structure validation, protein structure validation, PDB, PDB_REDO

Background

The 3D structure of proteins depends on their amino acid sequence [1], but cannot be predicted based solely on that sequence, except for relatively small proteins [2]. As the structure of a molecule cannot be observed directly, a model of the structure must be constructed using experimental data. These data can be obtained through different methods, such as X-ray crystallography, NMR spectroscopy or electron microscopy. However, none of these methods allows direct calculation of the structure from the data. In X-ray crystallography, the most applied method, the crystallographic diffraction data are used to construct a three-dimensional grid that represents the probability of electrons to be present in specific positions in space, the so-called electron density (ED) map. The ED shows an average between many (typically 10^{13} and 10^{15}) molecules arranged in a periodic fashion in crystals, and is also an average over the time of the X-ray experiment [3]. This ED is then interpreted to construct an atomic model of the structure. The model is just a representation of the crystallographic data and other known information about the structure, like the sequence, bond lengths and angles. Different models, such as the thousands of models in the Protein Data Bank (PDB) [4], represent the experimental data within varying degrees of reliability, and the quality of quantity of experimental data (for example the resolution limit of the diffracted X-rays) varies also a lot.

Due to the interpretation step during modelling, which is inevitably subjective [5, 6], it is very important to see if a model fits reasonably to the ED that was used to construct it, to ensure its reliable. For drug discovery and design purposes, the model quality of the protein binding sites and of the ligands

bound to them are of particular interest, while the overall model quality or the quality of part of the model outside the binding site, are not directly relevant.

A good way to assess how well a subset of atomic coordinates fits the experimental electron density, is the Real Space R-value (RSR) [7], which has been recommended by the X-ray Validation Task Force of the Worldwide PDB [8, 9]. The RSR measures a similarity score between the $2mFo-DFc$ and the DFc maps. The real-space correlation coefficient (RSCC) [6] is another well-established measure of model fit to the experimental data. The use of the ED to validate the model will not catch all possible problems in the model [10], but they can show whether the model fits the data it was created from.

VHELIBS aims to enable non-crystallographers and users with little or no crystallographic knowledge, to easily validate protein structures, before using them in drug discovery and development. To that end, it features a Graphical User Interface (GUI) with carefully chosen default values, valid for most situations, but also allowing easy tuning of parameters for more advanced users. A tool named Twilight [11, 12] has been recently published to evaluate ligand density. However, while VHELIBS focuses on assessing both the ligands and binding sites to aid model evaluation for drug discovery purposes, Twilight is ligand-centric and focuses on highlighting poorly modelled ligands. VHELIBS also enables the user to choose between the models from either the PDB [4, 13] or the PDB_REDO [14] databanks. Using PDB_REDO as the data source can have substantial benefits over using the PDB. PDB_REDO changes models both by re-refinement, incorporating advances in crystallographic methods since the original structure model (the PDB entry) was constructed and with limited rebuilding, mainly of residue side chains [15], improving the fit of models to the ED [16].

Implementation

VHELIBS validates the binding site and ligand against the ED, in a semi-automatic way, classifying them based on a scoring schema as 'good', 'bad' or 'dubious'. This score is calculated taking several parameters into account (RSR, RSCC, and average occupancy by default, but more can be used). After performing the automatic analysis and classification of a target's binding site and ligand, it then enables the user to graphically review and compare them with their ED in order to make it easier to properly classify any 'dubious'-labelled structure or re-classify any other structure based on actual visual inspection of the ED with the model.

VHELIBS is mainly implemented using Python under Jython [17], with some critical parts implemented in Java. It uses Jmol [18] for the 3D visualization of models and EDs. Electron density maps are retrieved from the EDS [19, 20] or from the PDB_REDO databank, which are updated weekly with new data from the PDB. Models are downloaded from either the PDB or PDB_REDO according to the user settings.

Description of the algorithm

VHELIBS takes as input a user-provided list of either PDB [13] or UniProtKB [21] codes. The codes in these lists can be entered directly from the GUI or be provided in a text file. UniProtKB codes are mapped to their corresponding PDB codes, so what is parsed in the end is a list of PDB codes.

For each of these PDB codes, statistical data are retrieved from the EDS or from the PDB_REDO, depending on the source of the models being analysed (i.e. EDS data for models downloaded from the PDB and PDB_REDO data for models downloaded from the PDB_REDO). Ligands bound with residues or molecules included in the 'blacklist' exclusion list (see below) with a bond length $< 2.1 \text{ \AA}$ are rejected. Those bound to molecules in the 'non-propagating' exclusion list (which can be modified by the user and by default contains mainly metal ions) are not rejected. The exclusion lists are composed of the most common solvent molecules and other non-ligand hetero compounds often found in PDB files, as well as some less common solvents and molecules which were found to have very simple binding sites (e.g. a binding site consisting of just 1-2 residues). We also incorporated the buffer molecules from Twilight's list [11, 12]. The exclusion list from BioLip [22] was also considered, but deemed too restrictive.

Once the ligands are determined, all the residues nearer than a specified distance (4.5 \AA by default) are considered part of the binding site of that ligand. Then, every ligand and binding site residue is given a score and classified by that score based on the following algorithm (see also Figure 1):

- For each residue and component of each ligand and each binding site, the initial score is defined to be 0.
- For each unmet user-specified condition, the score is increased by 1. The user specified conditions are the value thresholds for several different properties of the model and the data (i.e. RSR, RSCC, occupancy-weighted B factor, R-free, resolution and residue average occupancy; the user may use only some of these properties).
- If the score remains 0, the ligand/residue is labelled as *Good*.

- If the score is greater than the user-defined tolerance value, the ligand/residue is labelled as *Bad*.
- If the score is between 0 and the user-defined tolerance value, the ligand/residue is labelled as *Dubious*.
- At the end of all evaluations, the binding site and the ligand (for ligands with more than 1 'residue', i.e. those composed by more than one hetero compound in the PDB file) are labelled according to the highest score of their components (i.e. a binding site with a *Bad* residue will be labelled as *Bad* regardless of how the rest of the residues are labelled, and a binding site can only be labelled as *Good* when all its residues are labelled as *Good*).

The results from this classification are saved to a CSV file (the results file) which can be opened by any major spreadsheet software and from there they can be filtered as desired (for good ligands, for good binding sites or for both). A file with a list of all the rejected PDB structures and ligands and the reason for the rejection is also generated next to the results file.

Up to this point, the automatic classification of ligands and binding sites is complete. Now the user can visually inspect the results in order to see whether a binding or ligand labelled as *Dubious* can actually be marked as *Good* or not. When doing so, the user is presented with an interface like the one showed in Figures 2 and 3. The visualization of the binding site, the ligand and coordinates to examine (dubious or bad residues and ligands) and their respective EDs, can be customised in several ways through the GUI, being able to change colours, styles and even the contour level and radius of the EDs. Thus, the default visualization settings provide VHELIBS' users with the appropriate frame to easily reclassify dubious residues and ligands either as good or bad:

- good binding site residues are showed by default in white and with a wireframe. style in order to show the context where the possible reclassification is evaluated.

- coordinates to exam for correctness are shown in ball and stick style and coloured according to their B-factor.
- ligand coordinates are shown in ball and stick style and coloured in magenta (but can be coloured according to their B-factor if they need to be examined).
- ED for coordinates to exam is shown in yellow.
- ED for the complete binding site can be added to the visualization (in cyan) if necessary.
- ED for the ligand can be added to the visualization (in red) if necessary.

Hence, with this visualization frame, the user has all the information he/she needs in order to decide, for instance, whether: (a) binding site dubious coordinates could be relevant or not for protein-ligand docking results (if the dubious coordinates face opposite to the ligand, it is reasonable to think that their correctness does not affect protein-ligand docking results); and (b) ligand coordinates that were classified as bad or dubious by the automatic analysis can be changed to good because its experimental pose is the only possible for its corresponding ED (this usually happens with non flexible rings that have only ED for some of their atoms). In the online documentation (<https://github.com/URVnutrigenomica-CTNS/VHELIBS/wiki>) [23] there is more information on this and some practical rules for guiding such evaluation.

VHELIBS can be used with different running conditions (i.e. with different *profiles*). The values of the default profiles [i.e. *Default (PDB)* and *Default (PDB_REDO)*] were chosen after careful visualization and comparison of models with their EDs, giving a default minimum RSCC of 0.9, a minimum average occupancy of 1.0, a maximum RSR of 0.4 and a maximum good RSR of 0.24 for PDB and 0.165 for PDB_REDO. The different RSR cut-offs for the PDB and PDB_REDO are the result of RSR being calculated using different software in the EDS (which uses MAPMAN [24]) and in PDB_REDO (which uses EDSTATS [25]). The third provided profile, *Iridium*, is based on the values used in the construction of the Iridium set [26]. This profile is only provided as an example of how easy it is to adapt VHELIBS to use other values found in

literature. Note however that VHELIBS will yield slightly different results to those in the Iridium set, because VHELIBS uses the EDs and statistical data from EDS or PDB_REDO, while the authors of the Iridium set calculate all the data using different software and different EDs.

Key features of VHELIBS

- Many different parameters can be used to filter good models, and their threshold values can be adjusted by the user. Contextual help informs the user about the meaning of the different parameters.
- VHELIBS comes with three “profiles”, and the user can create custom profiles and export them for further use or sharing.
- Ability to work with an unlimited amount of PDB codes (which can be read from a list).
- Ability to choose between models from PDB_REDO or from the PDB.
- Ability to work with an unlimited amount of UniProtKB [27] accession numbers and names (which can be read from a list). In that situation, all the PDB codes included at each UniProtKB entry in the list are analysed by VHELIBS.
- VHELIBS runs in the Java Virtual Machine, and thus can run on any system with a recent version of a Java Runtime Environment, which is very likely already installed in user machines.
- VHELIBS consists of a single jar file, needing no installation. There are no dependencies other than Java.
- The user can load a results file from previous analysis and review the structures to see the ED fitting and then reclassify the model if needed. The resulting file from this can be opened again to resume, correct or just review the model with its ED, so the user can let a huge analysis run over lunch or overnight and then review the results later at any time.

- A user does not need to be familiar with any other software (although familiarity with Jmol [18] will help the user to make custom sophisticated views).

PDB_REDO changes to support VHELIBS

The PDB_REDO databank was upgraded to have per-residue RSR and RSCC values and downloadable EDs in the CCP4 [28] format for each entry. These ready-made maps not only make electron density visualisation possible in VHELIBS, but also in PyMOL [29] (a new plugin is available via the PDB_REDO website).

To assess how much of the previously observed model improvement in PDB_REDO is applicable to ligands and their binding pocket, we implemented two new ligand validation routines in the PDB_REDO pipeline: **(1)** EDSTATS [25] calculates the fit of the ligand with the ED; and **(2)** YASARA [30] calculates the heat of formation of the ligand (which is used as a measure of geometric quality) and interactions of the ligand with the binding pocket. The interactions measured in YASARA include the number of atomic clashes (bumps), number and total energy of hydrogen bonds, and the number and strength of hydrophobic contacts, π - π interactions, and cation- π interactions. The strengths of hydrophobic contacts, π - π interactions, and cation- π interactions are based on knowledge-based potentials [31] in which each individual interaction has a score between 0 and 1.

Results and Discussion

We performed an analysis of the ligand quality scores in the PDB and PDB_REDO, for more than 16,500 ligands (compounds described by the PDB as ‘non-polymer’; not chemically linked to the protein; with common crystallisation additives, such as sulphate and glycerol, excluded) in more than 5900 structures and the results are summarised in Table 1. The results show that ligands in PDB_REDO are better in terms of fit to the ED (better RSR and RSCC) and have more favourable geometry (lower heat of formation). Although the interactions with binding sites improve, the changes are very

small, except for the reduction in atomic clashes. This is to be expected, as ligand binding sites are typically the most important part of a structure model and much attention is paid to ensure the model is correct in that area. Nevertheless, in individual cases the improvement can be big enough to change a dubious ligand in a bad binding site to a good ligand in a good binding site (Figure 4).

All ligands and binding sites present in both the EDS and the PDB_REDO databanks were analysed, using the appropriate default profiles [*Default (PDB)* and *Default (PDB_REDO)*]. The results are summarised in Table 2 (for the binding sites) and Table 3 (for the ligands). In the case of the binding sites, the *Good* binding sites in the EDS are the 19.26%, while in the PDB_REDO they are the 35.92%, although only the 67% of the *Good* binding sites in the EDS are classified as *Good* for the PDB_REDO, even having some of them classified as *Bad*. In the case of the ligands, however, the improvement in classification from the PDB_REDO is far more significant: *Good* ligands go from 31.19 % from the EDS to a 63.66% from the PDB_REDO, having most of the *Good* ligands from EDS still classified as *Good* from PDB_REDO (94.54 %), and dramatically reducing *Bad* ligands from a 43% from EDS to a 3.64% from PDB_REDO, having most of these *Bad* ligands from EDS classified as *Good* from the PDB_REDO. Interestingly, our results suggest that, by default, a typical VHELIBS user should choose the *Default (PDB_REDO)* profile instead of the *Default (PDB)* one. From the 16830 binding sites that are labelled as *Good* by either of the default profiles, 85% of them are identified by the *Default (PDB_REDO)* profile [in contrast with only 45.58% being identified by the *Default (PDB)* profile]. This is even more remarkable when the ligands are considered: from the 26028 ligands labelled as *Good* by either of the default profiles, 97.4% of them are identified by the *Default (PDB_REDO)* profile and only 47.7% are identified by the *Default (PDB)*.

To demonstrate how VHELIBS can be used, we chose as a test case the human Dipeptidyl peptidase 4 (DPP-IV). We first used the corresponding UniProtKB name, DPP4_HUMAN, with the *Default (PDB_REDO)* profile. There are 74 different PDB structures listed in the UniProtKB entry for this protein. The automatic analysis of all of these structures took an average of 2 min. and 0.43

s. on an AMD FX-8150 machine running Ubuntu 12.04.1 LTS amd64 and Java (OpenJDK) 1.6.0_24, with some of the time being spent downloading data from the PDB_REDO (with cached PDB_REDO data, and thus without downloading it, the average is 1 min. 15.78 s.).

Out of the original 74 PDB structures, 10 were rejected because there was no PDB_REDO data available for them (1J2E, 1NU6, 1NU8, 1R9M, 1R9N, 1RWQ, 1WCY, 2BUB, 2JID and 2QKY). This mostly happens when a PDB entry lacks experimental X-ray reflection data, which is the case for the ten structures listed. From the remaining 64 structures, 44 had no ligands, resulting to 20 structures. These 20 PDB_REDO models showed 450 possible ligand-binding site pairs, of which 9 were rejected because the ligand was covalently bound to a residue, and 366 were rejected because the ligand was either blacklisted or covalently bound to a blacklisted ligand. Most of these rejected ligand-binding sites include molecules such as SO4 that are marked as hetero compounds by the PDB, covalently bound ligands (e.g. mannose/MAN in 2BGN), or metal ions (e.g. sodium or mercury) that are not usually used for drug discovery purposes. The valid ligand-binding site pairs were 75. Of these, 55 were labelled as good ligands, 57 as good binding sites and 43 as good ligand and binding site (Table 4).

With 55 good ligands and 57 good binding sites (43 of them being good binding sites with good ligands) there should be enough *good* structures for most use cases, so it would not be necessary to review the *dubious* ones. However, if that were not the case, the user can review *dubious* cases to see if they could be good enough for the specific purposes. Figure 2 show one example of a good ligand with a dubious binding site whereas Figure 3 shows a dubious ligand with a bad binding site. The user can also review the good structures if looking for false positives, or review the bad ones in the hope of finding good enough structures there (which is very unlikely using the default profiles).

Conclusions

Currently there is no other tool to easily check model to ED fitting for binding sites and ligands, and available alternatives need a lot of scripting or console commands for each structure.

There are several use cases where VHELIBS can prove very helpful:

- When choosing structures to use for a protein-ligand docking: with VHELIBS the user can choose the structures with the best modelled binding sites.
- For choosing structures where both the binding site and the ligand are well modelled, in order to validate the performance of different docking results. This could make it possible to obtain a new gold standard for protein/ligand complexes that can be used for the validation of docking software and that could be significantly larger and more diverse than those being currently used (i.e. the Astex Diverse Set [32] and the Iridium set [26]).
- For choosing structures where both the binding site and the ligand are well modelled, in order to obtain reliable structure-based pharmacophores that pick the intermolecular interactions that are relevant for modulating the target bioactivity. This is important in drug-discovery workflows for finding new molecules with similar activity to the co-crystallised ligand.
- To obtain well modelled ligand coordinates in order to evaluate the performance of 3D conformation generator software that claims to be able to generate bioactive conformations.

Our study allows to conclude also that, in general, binding site and ligand coordinates derived from PDB_REDO structures are more reliable than those obtained directly from the PDB and, therefore, highlights the contribution of the PDB_REDO database to the drug-discovery and development community.

Availability and Requirements

- **Project name:** VHELIBS (Validations Helper for Ligands and Binding Sites)
- **Project home page:** <http://urvnutrigenomica-ctns.github.com/VHELIBS/>
- **Operating System(s):** Platform independent
- **Programming language:** Python, Java
- **Other requirements:** Java 6.0 or newer, internet connection.
- **License:** GNU AGPL v3
- **Any restrictions to use by non-academics:** None other than those specified by the license (same as for academics).

List of abbreviations

ED: Electron Density

PDB: Protein Data Bank

GUI: Graphical User Interface

RSR: Real Space Residual

RSCC: Real Space Correlation Coefficient

DPP-IV: Dipeptidyl peptidase 4

Competing interests

The author(s) declare that they have no competing interests

Author's contributions

ACM, SGV, and GP designed the software and prepared the manuscript. RPJ and AP advised on default parameters and enabled the use of PDB_REDO in VHELIBS and also contributed to the manuscript. Testing and feedback for new ideas and GUI design was done by MJO, RPJ, CV, MM, MJS, AAA and LA. The software implementation was done by ACM with the help of MJO.

Acknowledgements

This manuscript has been edited by American Journal Experts.

Funding: This work was supported by the Ministerio de Educación y Ciencia of the Spanish Government [AGL2008-00387 and AGL2011-25831] and the ACCIÓ program from the Generalitat de Catalunya [TECCT11-1-0012].

We acknowledge support from the Generalitat de Catalunya through grant XRQTC.

We also acknowledge Professor Robert Hanson from the St. Olaf College, for his support in questions regarding Jmol and Ed Pozharski for writing the initial PyMOL plugin.

References

1. Anfinsen CB: Principles that govern the folding of protein chains. *Science* 1973, 181:223–30.
2. Bradley P, Misura KMS, Baker D: Toward high-resolution de novo structure prediction for small proteins. *Science* 2005, 309:1868–71.
3. Rhodes G, Cooper J: Model and Molecule. In *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Academic Press; 2006:1–5.
4. Berman H, Henrick K, Nakamura H: Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 2003, 10:980.
5. Dauter Z, Weiss MS, Einspahr H, Baker EN: Expectation bias and information content. *Acta Crystallographica Section D Biological Crystallography* 2013, 69:141–141.
6. Brändén C-I, Alwyn Jones T: Between objectivity and subjectivity. *Nature* 1990, 343:687–689.
7. Jones TA, Zou JY, Cowan SW, Kjeldgaard M: Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallographica Section A Foundations of Crystallography* 1991, 47:110–119.
8. Read RJ, Adams PD, Arendall WB, Brunger AT, Emsley P, Joosten RP,

- Kleywegt GJ, Krissinel EB, Lütke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G, Zwart PH: A new generation of crystallographic validation tools for the protein data bank. *Structure* 2011, 19:1395–412.
9. Gore S, Velankar S, Kleywegt GJ: Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 2012, 68:478–83.
10. Richardson JS, Richardson DC: Studying and polishing the PDB's macromolecules. *Biopolymers* 2012:n/a–n/a.
11. Pozharski E, Weichenberger CX, Rupp B: Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Crystallogr D Biol Crystallogr* 2013, 69:150–67.
12. Weichenberger CX, Pozharski E, Rupp B: Visualizing ligand molecules in twilight electron density. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2013, 69:195–200.
13. Berman HM: The Protein Data Bank. *Nucleic Acids Research* 2000, 28:235–242.
14. Joosten RP, Vriend G: PDB improvement starts with data deposition. *Science* 2007, 317:195–6.
15. Joosten RP, Joosten K, Cohen SX, Vriend G, Perrakis A: Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics* 2011, 27:3392–8.
16. Joosten RP, Joosten K, Murshudov GN, Perrakis A: PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallographica Section D* 2012, 68:484–496.
17. The Jython Project [<http://www.jython.org/>].
18. Hanson RM: Jmol – a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography* 2010, 43:1250–1260.
19. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA: The Uppsala Electron-Density Server. *Acta Crystallogr D Biol Crystallogr* 2004, 60:2240–9.
20. EDS - Uppsala Electron Density Server [<http://eds.bmc.uu.se/eds/>].
21. Magrane M: UniProt Knowledgebase: a hub of integrated protein data.

Database (Oxford) 2011, 2011:bar009.

22. Yang J, Roy A, Zhang Y: BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* 2013, 41:D1096–103.

23. VHELIBS Online Documentation [<https://github.com/URVnutrigenomica-CTNS/VHELIBS/wiki>].

24. Kleywegt GJ, Jones TA: xdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Crystallogr D Biol Crystallogr* 1996, 52:826–8.

25. Tickle IJ: Statistical quality indicators for electron-density maps. *Acta Crystallogr D Biol Crystallogr* 2012, 68:454–67.

26. Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD: Essential considerations for using protein-ligand structures in drug discovery. *Drug Discov Today* 2012, 17:1270–1281.

27. UniProtKB [<http://www.uniprot.org/help/uniprotkb>].

28. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS: Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 2011, 67:235–42.

29. Schrödinger L: The PyMOL Molecular Graphics System. 2010.

30. Krieger E, Koraimann G, Vriend G: Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. *Proteins* 2002, 47:393–402.

31. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K: Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* 2009, 77 Suppl 9:114–22.

32. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW: Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* 2007, 50:726–41.

33. Edmondson SD, Mastracchio A, Mathvink RJ, He J, Harper B, Park Y-J, Beconi M, Di Salvo J, Eiermann GJ, He H, Leiting B, Leone JF, Levorse DA,

Lyons K, Patel RA, Patel SB, Petrov A, Scapin G, Shang J, Roy RS, Smith A, Wu JK, Xu S, Zhu B, Thornberry NA, Weber AE: (2S,3S)-3-Amino-4-(3,3-difluoropyrrolidin-1-yl)-N,N-dimethyl-4-oxo-2-(4-[1,2,4]triazolo[1,5-a]pyridin-6-ylphenyl)butanamide: a selective alpha-amino amide dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes. *J Med Chem* 2006, 49:3614–27.

34. RCSB Protein Data Bank - RCSB PDB - 3Q8W Structure Summary [<http://www.rcsb.org/pdb/explore/explore.do?structureId=3Q8W>].

35. Vos S, Parry RJ, Burns MR, De Jersey J, Martin JL: Structures of free and complexed forms of Escherichia coli xanthine-guanine phosphoribosyltransferase. *J Mol Biol* 1998, 282:875–89.

Figures

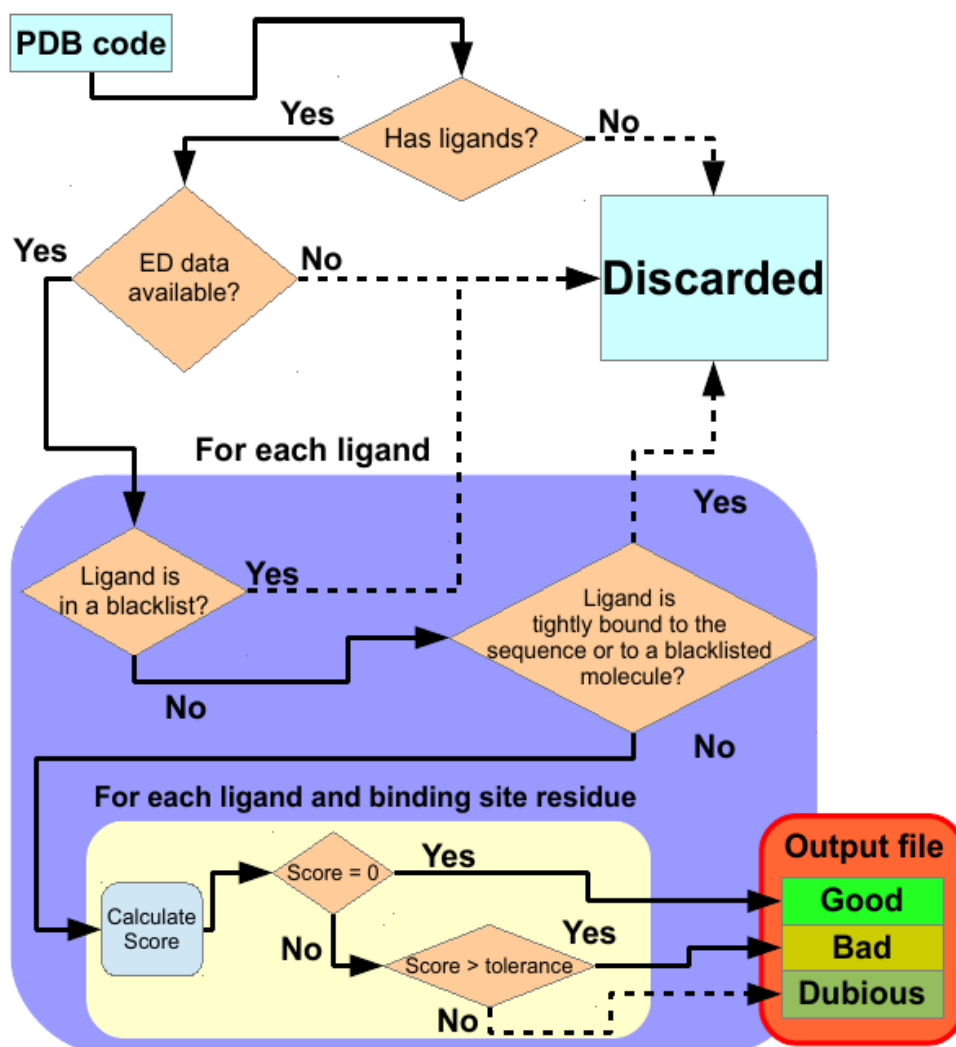


Figure 1: *Automatic ligand and binding site classification*; This diagram shows the process by which the ligands and binding sites of each PDB/PDB_REDO model are classified based on how well does the model fit the ED.

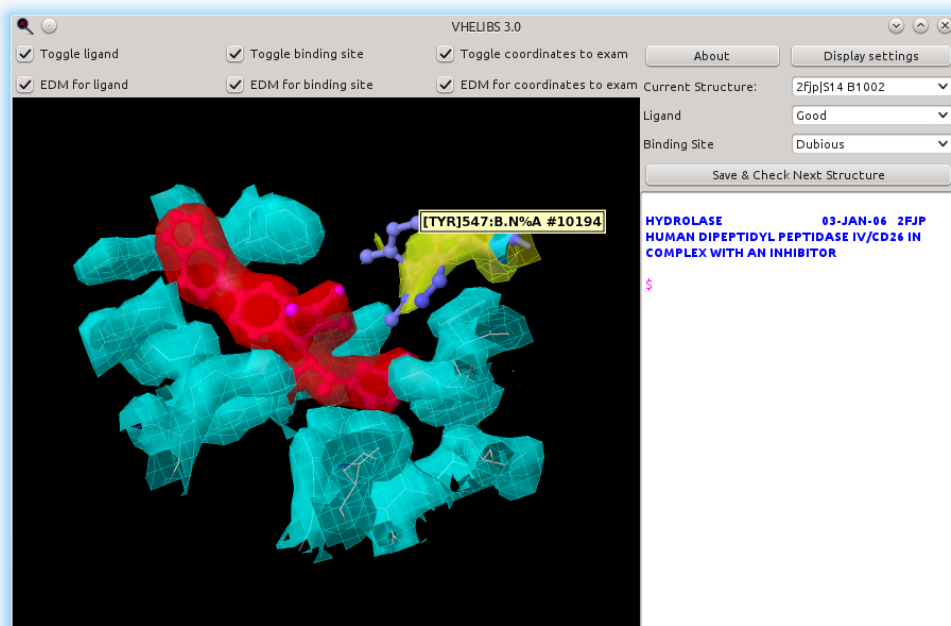


Figure 2: Example of a good ligand with a dubious binding site; Here we can see a ligand (S14 B1002 in PDB entry 2FJP [33]) and its binding site, from the results from the analysis of DPP4_HUMAN using the *Default* (PDB) profile. The only dubious residue from the binding site is the one with the yellow ED and represented by ball and stick, coloured by temperature factor.

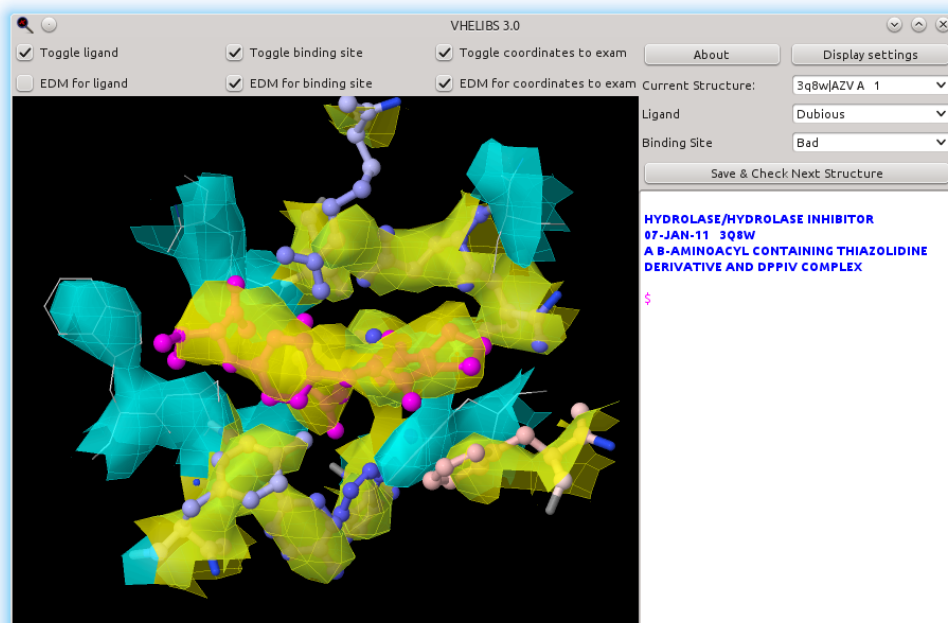


Figure 3: *Example of a dubious ligand with a bad binding site;* Here we can see a ligand (AZV A 1 in PDB entry 3Q8W [34]) and its binding site, from the same analysis as in Figure 2. As can be seen, some residues from this binding site hardly fit their ED (in yellow). The ligand mostly fits its ED, but it still has some discrepancies.

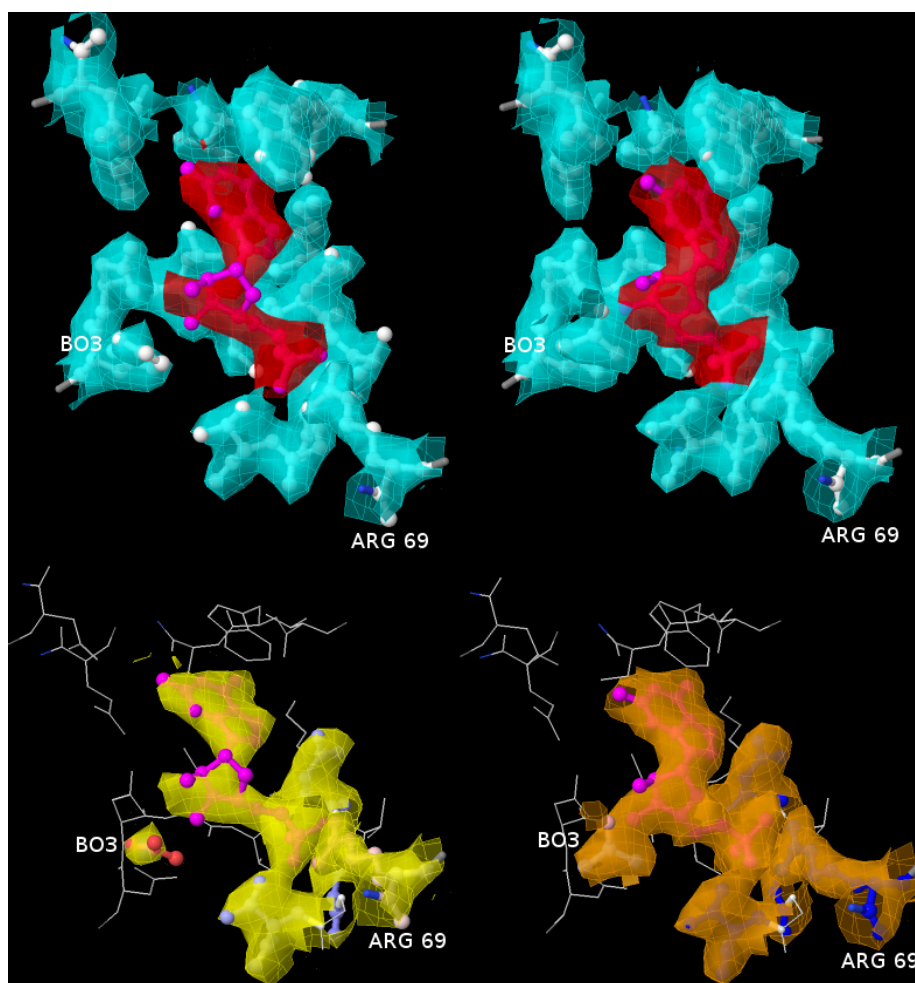


Figure 4: The guanosine-5'-monophosphate binding site in chain C of PDB entry 1A97 [35] is an example of a ligand and binding site flagged respectively as dubious and bad in the PDB (left panel: upper with cyan ED for the binding site and red ED for the ligand; lower with default view: yellow ED for dubious and bad residues), but as good in PDB_REDO (right panel: upper with cyan ED for the binding site and red ED for the ligand; lower with previously bad or dubious residues with orange ED). The RSR and RSCC of the ligand improve from 0.154 to 0.065 and from 0.86 to 0.97, respectively. Two extra hydrogen bonds are introduced, improving the total hydrogen bonding energy from -157 kJ/mol to -199 kJ/mol. The all atom root mean square deviation of the ligand is

0.6 Å. Of the residues in the binding site, arginine 69 and the boric acid molecule improve most significantly in terms of fit to the ED.

Tables

Table 1: Average validation scores for ligands in PDB and PDB_REDO

| Validation score ^a | PDB average ^b | PDB_REDO average ^b |
|---|--------------------------|-------------------------------|
| RSR ^c | 0.120 | 0.104 |
| RSCC ^c | 0.90 | 0.92 |
| Heat of formation (kJ/mol) ^d | -1011 | -1067 |
| Hydrogen bonding energy (kJ/mol) ^d | -57.7 | -58.8 |
| Hydrophobic contact strength ^{d,e} | 16.20 | 16.43 |
| π - π interaction strength ^{d,e} | 1.26 | 1.28 |
| cation- π interaction strength ^{d,e} | 1.17 | 1.19 |
| Number of atomic clashes ^d | 9.1 | 7.9 |

^a Lower is better for RSR, heat of formation, hydrogen bonding energy and number of atomic clashes. Higher is better for RSCC, hydrophobic contact strength, π - π interaction strength and cation- π interaction strength.

^b Overage over 16904 ligands (13703 for heat of formation) in 5932 structure models.

^c Calculated using EDSTATS [25]

^d Calculated using YASARA [30] using the atomic coordinates as-is.

Strained ligand conformations give high values.

^e The average reported is the average sum of all interactions for a single ligand.

Table 2. All binding sites present in both PDB and PDB_REDO were analysed. In this table it can be seen how were they classified when coming from the EDS or from the PDB_REDO databank .

| | | PDB_REDO | | | |
|-----|---------|----------|------|---------|-------|
| | | good | bad | dubious | |
| EDS | good | 5145 | 1600 | 926 | 7671 |
| | bad | 5500 | 3727 | 8395 | 17622 |
| | dubious | 3659 | 2953 | 7915 | 14527 |
| | | 14304 | 8280 | 17236 | 39820 |

Table 3. All ligands present in both PDB and PDB_REDO were analysed. In this table it can be seen how were they classified when coming from the EDS or from the PDB_REDO databank.

| | | PDB_REDO | | | |
|-----|---------|----------|------|---------|-------|
| | | good | bad | dubious | |
| EDS | good | 11741 | 16 | 662 | 12419 |
| | bad | 9819 | 1206 | 6098 | 17123 |
| | dubious | 3790 | 229 | 6259 | 10278 |
| | | 25350 | 1451 | 17236 | 39820 |

Table 4. Number of complexes classified as good, bad or dubious after applying VHELIBS to 75 ligand/DPP-IV binding site complexes using the Default (PDB_REDO) profile.

| | | binding site | | | |
|--------|---------|--------------|-----|---------|----|
| | | good | bad | dubious | |
| ligand | good | 43 | 0 | 12 | 55 |
| | bad | 0 | 0 | 0 | 0 |
| | dubious | 14 | 0 | 6 | 20 |
| | | 57 | 0 | 18 | 75 |

Tools for *In Silico* Target Fishing

Adrià Cereto-Massagué¹, María José Ojeda¹, Cristina Valls¹, Miquel Mulero¹,
Gerard Pujadas^{1,2}, Santiago Garcia-Vallve^{1,2*}

¹Group of Cheminformatics & Nutrition. Biochemistry and Biotechnology Department, Universitat Rovira i Virgili (URV), Tarragona, Catalonia, Spain

² Centre Tecnològic de Nutrició i Salut (CTNS), TECNIO, Reus, Spain

* Corresponding author: santi.garcia-vallve@urv.cat

Abstract

Computational target fishing methods are designed to identify the most probable target of a query molecule. This process may allow the prediction of the bioactivity of a compound, the identification of the mode of action of known drugs, the detection of drug polypharmacology, drug repositioning or the prediction of the adverse effects of a compound. The large amount of information regarding the bioactivity of thousands of small molecules now allows the development of these types of methods. In recent years, we have witnessed the emergence of many methods for *in silico* target fishing. Most of these methods are based on the similarity principle, i.e., that similar molecules might bind to the same targets and have similar bioactivities. However, the difficult validation of target fishing methods hinders comparisons of the

performance of each method. In this review, we describe the different methods developed for target prediction, the bioactivity databases most frequently used by these methods, and the publicly available programs and servers that enable non-specialist users to obtain these types of predictions. It is expected that target prediction will have a large impact on drug development and on the functional food industry.

Highlights

- In recent years, a great number of methods for target prediction or drug repositioning have been developed.
- Most methods rely upon similarity with molecules whose bioactivity is known and other information for target prediction.
- The difficulties in validating predictions hinder comparisons of the performance of different methods.
- Target fishing methods could have a large impact on drug research and functional food industry.

Keywords

Computational Target Fishing, Reverse Screening, Drug Repositioning, Polypharmacology, Drug Research, Functional Foods

1. Introduction

In contrast to virtual screening, which is used to search large libraries of compounds for molecules that are most likely to bind a specific target, the aim of reverse screening, also known as *in silico* or computational target fishing [1,2] or reverse pharmacognosy [3], is to identify the most likely targets of a query molecule. This approach allows the prediction of the bioactivity of the query molecule or its mechanism of action. In addition, these techniques can be used to predict the adverse effects of a compound [4,5], to detect drug polypharmacology [6–8], or to reposition drugs [7,9–13].

Known drugs have, on average, six molecular targets on which they exhibit activity [14]. Polypharmacology, the ability of small molecules to interact with multiple proteins, is of particular interest for rationally designing more effective and less toxic drugs. Drug repositioning, the process of finding new uses for known drugs, is a promising way to explore alternative indications for existing drugs [13]. Because the successful launch of a single new drug is estimated to cost approximately U.S. \$800 million and takes a staggering 15 years, and because very few compounds that start a clinical trial emerge to the market [10], finding new uses for old drugs could be economically advantageous.

Taking into account that several databases, such as ChEMBL, contain millions of molecules and information about their bioactivity, it is now becoming feasible to merge the known “chemical space” and “biological space” into models that will enable us to generate biological “spectra” to predict the phenotypic activity of new molecules based on their chemical structures and the known bioactivities of structurally similar compounds [15]. Although the current methods of virtual screening could be successfully adopted for target

fishing, the differences in the general tasks of these methods justify the independent development of new *in silico* techniques for target fishing.

2. Computational methods for target fishing

Various computational methods have been developed to predict the molecular targets of a compound [1,16]. These methods were initially classified into four groups: chemical similarity searching, data mining/machine learning, panel docking, and the analysis of bioactivity spectra [16,17]. Recently, other classes, such as protein-structure-based methods, have been proposed [18]. Below, we summarize the main characteristics of some of these methods.

2.1 Molecular similarity methods

This section describes chemical similarity methods and shape-based similarity methods. The simplest methods for target prediction are based on chemical similarity and the use of current knowledge about the bioactivity of millions of small molecules. These methods are based on the “chemical similarity principle,” which states that similar molecules are likely to have similar properties [19,20]. Thus, the targets of a molecule can be predicted by identifying proteins with known ligands that are highly similar to the query molecule [16]. The advantage of these methods is that they only require the computation of the similarity between compounds [19,21]. An outline of a chemical similarity method is shown in Figure 1. In this method, a small molecule is represented as a chemical fingerprint. Fingerprints are a way of encoding the structure of a molecule. The most common type of fingerprint is a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule. The interested reader is referred to [22] for a review about fingerprints. To compare the fingerprints of two molecules, the

Tanimoto coefficient or any other similarity criterion can be used. The more similar two compounds are, the closer the Tanimoto coefficient will be to 1. Several databases describing the bioactivities of thousands or millions of small molecules or the activities of known drugs can be used for target prediction (see Table 1 and reference [1]).

Keiser et al. [23] used a similarity ensemble approach to compare protein targets by the 2D similarity of the ligands that they are known to bind. The authors screened a dataset of 3,665 drugs, including drugs approved by the FDA and investigational drugs, against a database of 65,241 ligands organized into 246 protein targets taken from the MDL Drug Data Report database. Their study revealed unanticipated associations between thousands of drugs and ligand sets [23]. Of the 30 most promising drug-target associations that were tested experimentally, 23 were confirmed, and 5 of the 23 were shown to be potent (<100nM) modulators of their predicted target [23]. Thus, their study demonstrated the power of using simple ligand-based similarity searches.

Because they can be calculated quickly, 2D fingerprints have been widely used for similarity searching in target fishing. However, 3D chemical descriptors can also be used [17], although calculating them is computationally more expensive. Because they contain more information, the predictions based on 3D fingerprints would be expected to be better than those based on 2D fingerprints. However, in some cases, methods that use 2D fingerprints outperform those methods that use 3D fingerprints in correct target prediction [24]. 3D descriptors work better in cases of low structural similarity [24].

A known limitation of chemical similarity approaches is that inactive compounds can sometimes exhibit good similarity with active molecules if they have been obtained by modifying an active compound at some key position that

was crucial for its interactions [25]. These inactive compounds can be false positive predictions of target fishing methods. In addition, in some cases, a large group of false negatives is also expected, because not all types of active compounds for a specific target have been identified.

Shape-based similarity methods use 3D shape comparisons between molecules, usually comparing the shape of the molecular volume, but other “shapes” can be compared, like the electrochemical surface. This can be done with software such as ROCS [26], Phase Shape [27], ESHAPE3D [28], PARAFIT [29], ShaEP [30] and USR [31] as some examples. Shape-based methods have the potential of detecting similarities between molecules with different atomic structures, thus making them specially useful for scaffold-hopping. Pharmacophores and some molecular fingerprints (like Spectrophores [32] and many pharmacophore-based fingerprints [33]) can also include 3D information [22,33]. All these 3D methods require ligand conformations. In many cases (where there is no known biologically active conformation for the molecule), a single low-energy conformer is used, although it can be biologically irrelevant. Another approach is to get the conformation of the molecules by aligning them to a known bioactive conformation of a known ligand. However, 2D fingerprint-based methods give better performance than 3D shape-based methods in virtual screenings [34]. In other cases, combining chemical and shape similarity measures significantly increases the target prediction accuracy [35].

After obtaining the highest similarity coefficient between a query compound and the compounds in an annotated database, it is important to assess the statistical significance of the similarity. Two structures are usually considered similar if the Tanimoto coefficient between them is higher than 0.85. However,

this value is not always reliable [36]. Keiser et al. [37] used an E-value computed from the 2D similarity with the set of ligands of a target. This E-value is derived from the statistics of similarity values with all ligands (above a certain threshold), and it indicates how likely it would be to find a molecule with a given average similarity to the set of ligands of a target. The SwissTargetPrediction server uses a probability derived from a cross-validation analysis to rank the targets and estimate the accuracy of the predictions [25].

2.2 Data mining and machine learning methods

One of the major challenges of an *in silico* target fishing method is to identify the biological consequences of the query molecule binding to its predicted targets. For this reason, more complex methods have been developed. Data mining and machine learning-based methods, also known as chemogenomic approaches, usually combine fingerprints and some type of machine learning approach, such as self-organizing maps [38], Bayesian classifiers [4], or network classification [39], to develop predictive models. These methods usually require the use of systematic nomenclature in the training set (normalized target names) [16,17] and depend on reliable training data sets [2]. Associations between target names and chemical sub-structures can be extracted automatically across target class sets with inductive machine learning. Chemical features correlated with specific target binding are then stored in the form of multiple-target models. The target fishing problem is thus one of compound classification on a grand scale involving thousands of individual target class models [17].

Bender et al. [4] used normalized side-effect annotations in the World Drug Index and a multcategory Bayes Model that employed ECFP4 fingerprints to

build a model for adverse drug reactions. On average, 90% of the adverse drug reactions observed with known, clinically used compounds were detected [4].

2.3 Protein structure-based methods

Other computational target fishing methods use the protein structure of the targets to predict novel bioactivities. Protein docking [40–42], pharmacophore searching [43], or protein–ligand interaction fingerprints can be used [18]. These methods are limited to targets with resolved structures. Docking a query molecule to a large group of x-ray resolved structures demands large computational power or an extraordinary amount of time. In addition, docking is not a very reliable way to investigate ligand-target interactions, as no statistically significant relationship exists between docking scores and ligand affinity [16]. Despite these limitations, specific docking programs and servers for target fishing, called inverse docking methods, have been developed (Table 2), in some cases reducing the computing time required and developing special scoring measures [18].

2.4 Methods based on analysis of bioactivity spectra

The activities of a compound across a series of biological readouts, such as gene expression profiles or protein microarrays, can also be viewed as molecular descriptors and used for target prediction [16,44,45]. These methods use experimental values of the query molecules and require a reference collection, such as the Connectivity Map [46], of gene expression profiles from cultured human cells treated with bioactive small molecules. Wang et al. [47] demonstrated that the on-target and off-target effects of a drug could be characterized by drug-induced *in vitro* genomic expression changes. The Mantra 2.0 web server [48] (Table 2) explores similarities between drug-

induced transcriptional profiles and represents this information as a network. Visual inspection of the neighboring drugs and communities helps to reveal modes of action and suggests new applications of known drugs [48].

A similar approach uses a disease gene expression signature, derived from the set of differentially expressed genes between a disease and a healthy control sample, that is compared to gene expression profiles of drugs. Drugs with gene expression patterns that are oppositional to the disease gene expression pattern represent putative novel therapeutic indications [11].

3. Validation of the methods

A fundamental issue when developing a novel method for predicting or classifying is validating the method. Validation allows the comparison of the performance of different methods. However, most of the articles describing a novel computational method for target fishing do not validate their results or compare their method with existing ones. Ideally, to compare the predictive capacity of different methods, the same dataset must be applied to all of the methods being compared. This dataset, used to test the performance of the predictive methods as a community standard, is usually called a benchmark dataset. For predicting drug–protein interactions, a benchmark dataset manually constructed by Yamanishi et al. [49] has frequently been used [42]. Recently, a benchmark dataset consisting of more than 155,000 ligand-protein pairs from 894 human protein targets has been proposed for future target prediction methods [15]. Although there are several web servers publicly available for target fishing (see Table 2), most of the developed methods are not available on-line or as stand-alone programs. This lack of availability is an important difficulty for comparing different methods.

One possible way of validating a target fishing method would be to check how often known targets fall within the best-scoring predicted ones in the output of a method [25]. Ideally, however, the known target-compound information must not be used by the predictive method; otherwise, the validation would be obvious. To obtain a more balanced dataset that better reflects the much larger number of non-interacting protein–ligand pairs, additional negative interactions must be included [25]. This requirement can be met by linking the molecules of the test set to randomly chosen targets [25]. Machine learning-based methods usually use cross-validation [50,51]. Cross-validation consists of defining a training set, which is used for training the method, and a test set (a group of compounds with known targets) that it is used to validate the method. However, cross-validation often overestimates model performance. Overfitting is another problem, which occurs when a model performs well on a training set and much worse on subsequent data.

Retrospective analysis has been used to validate some of the computational target fishing methods [25]. An example of retrospective analysis would be using an initial version of a database to train or create a method and then using molecules that have been added to a newer version of the same database to test the method. This strategy cannot be used to compare the performance of different methods. Gottlieb and coworkers [52] used 2,552 unique drug–disease associations that were being investigated in clinical trials to validate their method. Twenty-seven percent of the associations were predicted by their method [52]. This approach is an interesting way to validate a target fishing method, although not all of the associations that are being investigated are true.

The best way to validate a predictive method is experimentally. Lounkine et al. [5] performed a large-scale prediction and testing of drug activity on side-effect

targets. More than 600 marketed drugs were computationally screened against a set of 73 protein targets, and approximately half of the positive predictions were subsequently confirmed experimentally [5]. Cheng et al. [39] used a supervised inference method to predict new drug-target interactions for 12,483 FDA-approved and experimental drug-target binary links. *In vitro* assays confirmed the novel targets of five old drugs [39]. Campillos et al. [53] used phenotypic side-effect similarities to infer whether two drugs shared a target. When their method was applied to marketed drugs, unexpected drug-drug relationships were discovered. Using *in vitro* binding assays, the authors experimentally validated 12 out of 20 of the unexpected drug-drug relationships [53].

4. Examples of target predictions

Although a lot of methods of target fishing have been developed, only a few of them have confirmed their predictions *in vitro* or *in vivo* and have showed their capacity for predicting unexpected new cross-target binding events. Most of these unexpected relationships have been found in the field of drug repositioning. Because the safety profiles of approved drugs are known, the development of alternative indications are cheaper and potentially faster [13]. Below we summarize some examples in this field:

Using an electrostatic and shape 3D similarity search of a database of approved drugs to a previously identified inhibitor of DNA methyltransferase, Olsalazine, an approved anti-inflammatory drug was predicted, and further characterized, as a novel DNA hypomethylating agent [54].

From gene expression measurements from 100 diseases and gene expression measurements on 164 drug compounds, Sirota and coworkers [55] developed a

computational approach to predict novel therapeutic indications on the basis of comprehensive testing of molecular signatures in drug-disease pairs. From their predictions, these authors experimentally validated the use of the antiulcer drug cimetidine as a candidate therapeutic in the treatment of lung adenocarcinoma, demonstrating its efficacy both *in vitro* and *in vivo* [55]. In a similar approach, from the comparison between data measuring gene expression in Inflammatory Bowel Disease (IBD) samples and gene expression from 164 small-molecule drug compounds, Dudley and coworkers [56] found that topiramate, an anticonvulsant drug not previously described to demonstrate efficacy for IBD or any related disorders of inflammation or the gastrointestinal tract, might serve as a therapeutic option for IBD in humans.

Using side-effect similarities and a network analysis, Campillos and coworkers [53] identified new unexpected drug targets. Rabeprazole, an antiulcer drug, and the nervous system drugs paroxetine and fluoxetine were found to inhibit the dopamine receptor DRD3 and to bind the serotonin receptor HTR1D [53].

Using a network-based inference method, Cheng et al. [39] predicted, and then confirmed *in vitro*, that montelukast, an agonist of cysteinyl leukotriene 1 receptor, is also a DPP-IV inhibitor, and that diclofenac, simvastatin, ketoconazole, and itraconazole show polypharmacological features on estrogen receptors.

Based on a structural similarity with pharmacophores of a known prostanoid TP receptor, Ting and Khasawneh [57] showed that glybenclamine, an antidiabetic drug, has antithrombotic activity in mouse models.

Using the MANTRA web-server, based on network theory and non-parametric statistics on gene expression data, Iorio et al. [44] correctly predicted the mode

of action for nine anticancer compounds. In addition, they were able to discover the unexpected similarity between cyclin-dependent kinase 2 inhibitors and Topoisomerase inhibitors [44].

Based on the chemical similarity between ligands, Keiser et al. [37] found the unexpected relationships between methadone, emetine and ioperamide with muscarinic M3, alpha2 adrenergic and neurokinin NK2 receptors, respectively.

Using the one-dimensional drug profile matching Kovacs et al. [58] found that nitazoxanide, an antiprotozoal agent that interfere with the electron transfer, is also a peroxisome proliferator-activated receptor agonist, showing that nitazoxanide lower fasting blood glucose levels and improve insulin sensitivity in type diabetic rats.

5. Conclusions

In recent years, a large number of computational target fishing methods have been developed. This abundance has been made possible by the availability of large libraries of information about the bioactivity of compounds and by advances in methodology. Understanding the biological mechanisms of current drugs and integrating this information with additional resources are essential steps for making more reliable predictions. However, efforts are needed to validate the results of different prediction methods. The creation of a benchmark dataset will enable a proper comparison of the performance of *in silico* target prediction methodologies. Identification of new targets for novel compounds or existing drugs and the prediction of adverse effects will facilitate drug discovery and the development of new ingredients for functional foods.

Acknowledgements

This manuscript was edited for English grammar and usage by American Journal Experts. This study was supported by grant AGL2011-25831/ALI from the Spanish Government and ACC1Ó program [TECCT11-1-0012] and grant XRQTC from ‘Generalitat de Catalunya’.

References

- [1] A. Koutsoukas, B. Simms, J. Kirchmair, P.J. Bond, A. V Whitmore, S. Zimmer, et al., From in silico target prediction to multi-target drug design: current databases, methods and applications., *J. Proteomics*. 74 (2011) 2554–2574.
- [2] L. Wang, X.-Q. Xie, Computational target fishing: what should chemogenomics researchers expect for the future of in silico drug design and discovery?, *Future Med. Chem.* 6 (2014) 247–249.
- [3] S. Blondeau, Q.T. Do, T. Scior, P. Bernard, L. Morin-Allory, Reverse pharmacognosy: another way to harness the generosity of nature., *Curr. Pharm. Des.* 16 (2010) 1682–1696.
- [4] A. Bender, J. Scheiber, M. Glick, J.W. Davies, K. Azzaoui, J. Hamon, et al., Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure., *ChemMedChem*. 2 (2007) 861–873. doi:10.1002/cmdc.200700026.

- [5] E. Lounkine, M.J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J.L. Jenkins, et al., Large-scale prediction and testing of drug activity on side-effect targets., *Nature*. 486 (2012) 361–367. doi:10.1038/nature11159.
- [6] A.S. Reddy, S. Zhang, Polypharmacology: drug discovery for the future., *Expert Rev. Clin. Pharmacol.* 6 (2013) 41–47. doi:10.1586/ecp.12.74.
- [7] J. Achenbach, P. Tiikkainen, L. Franke, E. Proschak, Computational tools for polypharmacology and repurposing., *Future Med. Chem.* 3 (2011) 961–968. doi:10.4155/fmc.11.62.
- [8] V.I. Pérez-Nueno, V. Venkatraman, L. Mavridis, D.W. Ritchie, Detecting drug promiscuity using Gaussian ensemble screening., *J. Chem. Inf. Model.* 52 (2012) 1948–1961.
- [9] T.T. Ashburn, K.B. Thor, Drug repositioning: identifying and developing new uses for existing drugs., *Nat. Rev. Drug Discov.* 3 (2004) 673–683. doi:10.1038/nrd1468.
- [10] C. Chong, D. Sullivan, New uses for old drugs, *Nature*. 448 (2007) 645–646.
- [11] J.T. Dudley, T. Deshpande, A.J. Butte, Exploiting drug-disease relationships for computational drug repositioning., *Brief. Bioinform.* 12 (2011) 303–311. doi:10.1093/bib/bbr013.
- [12] V.I. Pérez-Nueno, A.S. Karaboga, M. Souchet, D.W. Ritchie, GES Polypharmacology Fingerprints: A Novel Approach for Drug Repositioning., *J. Chem. Inf. Model.* 54 (2014) 720–734. doi:10.1021/ci4006723.
- [13] Z. Liu, H. Fang, K. Reagan, X. Xu, D.L. Mendrick, W. Slikker, et al., In silico drug repositioning: what we need to know., *Drug Discov. Today*. 18 (2013) 110–115. doi:10.1016/j.drudis.2012.08.005.

- [14] J. Mestres, E. Gregori-Puigjané, S. Valverde, R. V Solé, The topology of drug-target interaction networks: implicit dependence on drug properties and target families., *Mol. Biosyst.* 5 (2009) 1051–1057. doi:10.1039/b905821b.
- [15] A. Koutsoukas, R. Lowe, Y. Kalantarmotamedi, H.Y. Mussa, W. Klaffke, J.B.O. Mitchell, et al., In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window., *J. Chem. Inf. Model.* 53 (2013) 1957–1966. doi:10.1021/ci300435j.
- [16] A. Bender, D.W. Young, J.L. Jenkins, M. Serrano, D. Mikhailov, P.A. Clemons, et al., Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint., *Comb. Chem. High Throughput Screen.* 10 (2007) 719–731. doi:10.2174/138620707782507313.
- [17] J.L. Jenkins, A. Bender, J.W. Davies, In silico target fishing: Predicting biological targets from chemical structure, *Drug Discov. Today Technol.* 3 (2006) 413–421. doi:10.1016/j.ddtec.2006.12.008.
- [18] K.T. Schomburg, S. Bietz, H. Briem, A.M. Henzler, S. Urbaczek, M. Rarey, Facing the challenges of structure-based target prediction by inverse virtual screening., *J. Chem. Inf. Model.* 54 (2014) 1676–1686. doi:10.1021/ci500130e.
- [19] M.A. Johnson, G.M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, New York, 1990.
- [20] J.C. Adams, M.J. Keiser, L. Basuino, H.F. Chambers, D.-S. Lee, O.G. Wiest, et al., A mapping of drug space from the viewpoint of small molecule metabolism., *PLoS Comput. Biol.* 5 (2009) e1000474. doi:10.1371/journal.pcbi.1000474.

- [21] B. Chen, K.J. McConnell, N. Wale, D.J. Wild, E.M. Gifford, Comparing bioassay response and similarity ensemble approaches to probing protein pharmacology., *Bioinformatics*. 27 (2011) 3044–3049. doi:10.1093/bioinformatics/btr506.
- [22] A. Cereto-Massagué, M.J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, Molecular fingerprint similarity search in virtual screening., *Methods*. (2014). doi:10.1016/j.ymeth.2014.08.005.
- [23] M.J. Keiser, V. Setola, J.J. Irwin, C. Laggner, A.I. Abbas, S.J. Hufeisen, et al., Predicting new molecular targets for known drugs., *Nature*. 462 (2009) 175–181.
- [24] J.H. Nettles, J.L. Jenkins, A. Bender, Z. Deng, J.W. Davies, M. Glick, Bridging chemical and biological space: “target fishing” using 2D and 3D molecular descriptors., *J. Med. Chem.* 49 (2006) 6802–6810. doi:10.1021/jm060902w.
- [25] D. Gfeller, A. Grosdidier, M. Wirth, A. Daina, O. Michielin, V. Zoete, SwissTargetPrediction: a web server for target prediction of bioactive small molecules., *Nucleic Acids Res.* 42 (2014) W32–W38. doi:10.1093/nar/gku293.
- [26] P.C.D. Hawkins, A.G. Skillman, A. Nicholls, Comparison of Shape-Matching and Docking as Virtual Screening Tools, *J. Med. Chem.* 50 (2007) 74–82. doi:10.1021/jm0603365.
- [27] G.M. Sastry, S.L. Dixon, W. Sherman, Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring., *J. Chem. Inf. Model.* 51 (2011) 2455–2466. doi:10.1021/ci2002704.
- [28] Chemical Computing Group Inc., Molecular Operating Environment (MOE), (2013).

- [29] J.-H. Lin, T. Clark, An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties., *J. Chem. Inf. Model.* 45 (2005) 1010–1016. doi:10.1021/ci050059v.
- [30] M.J. Vainio, J.S. Puranen, M.S. Johnson, ShaEP: molecular overlay based on shape and electrostatic potential., *J. Chem. Inf. Model.* 49 (2009) 492–502. doi:10.1021/ci800315d.
- [31] P.J. Ballester, W.G. Richards, Ultrafast shape recognition to search compound databases for similar molecular shapes., *J. Comput. Chem.* 28 (2007) 1711–1723. doi:10.1002/jcc.20681.
- [32] SpectrophoresTM — Open Babel v2.3.1 documentation, (n.d.). <http://openbabel.org/docs/dev/Fingerprints/spectrophore.html> (accessed September 18, 2014).
- [33] M.J. McGregor, S.M. Muskal, Pharmacophore fingerprinting. 2. Application to primary library design., *J. Chem. Inf. Comput. Sci.* 40 (1999) 117–125.
- [34] V. Venkatraman, V.I. Pérez-Nueno, L. Mavridis, D.W. Ritchie, Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods., *J. Chem. Inf. Model.* 50 (2010) 2079–2093.
- [35] D. Gfeller, O. Michielin, V. Zoete, Shaping the interaction landscape of bioactive molecules., *Bioinformatics.* 29 (2013) 3073–3079. doi:10.1093/bioinformatics/btt540.
- [36] G. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, Molecular similarity in medicinal chemistry., *J. Med. Chem.* 57 (2014) 3186–3204. doi:10.1021/jm401411z.

- [37] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, B.K. Shoichet, Relating protein pharmacology by ligand chemistry., *Nat. Biotechnol.* 25 (2007) 197–206. doi:10.1038/nbt1284.
- [38] D. Reker, T. Rodrigues, P. Schneider, G. Schneider, Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus., *Proc. Natl. Acad. Sci. U. S. A.* 111 (2014) 4067–4072. doi:10.1073/pnas.1320001111.
- [39] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, et al., Prediction of drug-target interactions and drug repositioning via network-based inference., *PLoS Comput. Biol.* 8 (2012) e1002503. doi:10.1371/journal.pcbi.1002503.
- [40] Z. Gao, H. Li, H. Zhang, X. Liu, L. Kang, X. Luo, et al., PDTD: a web-accessible protein database for drug target identification., *BMC Bioinformatics.* 9 (2008) 104. doi:10.1186/1471-2105-9-104.
- [41] H. Luo, J. Chen, L. Shi, M. Mikailov, H. Zhu, K. Wang, et al., DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome., *Nucleic Acids Res.* 39 (2011) W492–W498. doi:10.1093/nar/gkr299.
- [42] Y.-C. Wang, C.-H. Zhang, N.-Y. Deng, Y. Wang, Kernel-based data fusion improves the drug–protein interaction prediction, *Comput. Biol. Chem.* 35 (2011) 353–362. doi:10.1016/j.compbiolchem.2011.10.003.
- [43] X. Liu, S. Ouyang, B. Yu, Y. Liu, K. Huang, J. Gong, et al., PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach., *Nucleic Acids Res.* 38 (2010) W609–W614. doi:10.1093/nar/gkq300.
- [44] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, et al., Discovery of drug mode of action and drug repositioning from

- transcriptional responses., *Proc. Natl. Acad. Sci. U. S. A.* 107 (2010) 14621–14626. doi:10.1073/pnas.1000138107.
- [45] D. Emig, A. Ivliev, O. Pustovalova, L. Lancashire, S. Bureeva, Y. Nikolsky, et al., Drug target prediction and repositioning using an integrated network-based approach., *PLoS One.* 8 (2013) e60618. doi:10.1371/journal.pone.0060618.
- [46] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, et al., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease., *Science.* 313 (2006) 1929–1935. doi:10.1126/science.1132939.
- [47] K. Wang, J. Sun, S. Zhou, C. Wan, S. Qin, C. Li, et al., Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity., *PLoS Comput. Biol.* 9 (2013) e1003315. doi:10.1371/journal.pcbi.1003315.
- [48] D. Carrella, F. Napolitano, R. Rispoli, M. Miglietta, A. Carissimo, L. Cuttillo, et al., Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis., *Bioinformatics.* 30 (2014) 1787–1788. doi:10.1093/bioinformatics/btu058.
- [49] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug-target interaction networks from the integration of chemical and genomic spaces., *Bioinformatics.* 24 (2008) i232–i240. doi:10.1093/bioinformatics/btn162.
- [50] T. van Laarhoven, S.B. Nabuurs, E. Marchiori, Gaussian interaction profile kernels for predicting drug-target interaction., *Bioinformatics.* 27 (2011) 3036–3043. doi:10.1093/bioinformatics/btr500.
- [51] X. Liu, I. Vogt, T. Haque, M. Campillos, HitPick: a web server for hit identification and target prediction of chemical screenings.,

- Bioinformatics. 29 (2013) 1910–1912.
doi:10.1093/bioinformatics/btt303.
- [52] A. Gottlieb, G.Y. Stein, E. Ruppin, R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine., *Mol. Syst. Biol.* 7 (2011) 496. doi:10.1038/msb.2011.26.
- [53] M. Campillos, M. Kuhn, A.-C. Gavin, L.J. Jensen, P. Bork, Drug target identification using side-effect similarity., *Science.* 321 (2008) 263–266. doi:10.1126/science.1158140.
- [54] O. Méndez-Lucio, J. Tran, J.L. Medina-Franco, N. Meurice, M. Muller, Toward drug repurposing in epigenetics: olsalazine as a hypomethylating compound active in a cellular context., *ChemMedChem.* 9 (2014) 560–565.
- [55] M. Sirota, J.T. Dudley, J. Kim, A.P. Chiang, A.A. Morgan, A. Sweet-Cordero, et al., Discovery and preclinical validation of drug indications using compendia of public gene expression data., *Sci. Transl. Med.* 3 (2011) 96ra77. doi:10.1126/scitranslmed.3001318.
- [56] J.T. Dudley, M. Sirota, M. Shenoy, R.K. Pai, S. Roedder, A.P. Chiang, et al., Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease., *Sci. Transl. Med.* 3 (2011) 96ra76.
- [57] H.J. Ting, F.T. Khasawneh, Glybenclamide: an antidiabetic with in vivo antithrombotic activity., *Eur. J. Pharmacol.* 649 (2010) 249–254.
- [58] D. Kovács, Z. Simon, P. Hári, A. Málnási-Csizmadia, C. Hegedús, L. Drimba, et al., Identification of PPAR γ ligands with One-dimensional Drug Profile Matching., *Drug Des. Devel. Ther.* 7 (2013) 917–928. doi:10.2147/DDDT.S47173.

- [59] E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant, PubChem: Integrated Platform of Small Molecules and Biological Activities, in: *Annu. Rep. Comput. Chem.*, American Chemical Society, Washington, DC, 2008: pp. 217–241. doi:10.1016/S1574-1400(08)00012-1.
- [60] N. Hecker, J. Ahmed, J. von Eichborn, M. Dunkel, K. Macha, A. Eckert, et al., SuperTarget goes quantitative: update on drug-target interactions., *Nucleic Acids Res.* 40 (2012) D1113–D1117. doi:10.1093/nar/gkr912.
- [61] M. Olah, M. Mracec, L. Ostopovici, R. Rad, A. Bora, N. Hadaruga, et al., WOMBAT: World of Molecular Bioactivity, in: *Chemoinformatics Drug Discov.*, 2005: pp. 221–239. doi:10.1002/3527603743.ch9.
- [62] J.J. Irwin, B.K. Shoichet, ZINC--a free database of commercially available compounds for virtual screening., *J. Chem. Inf. Model.* 45 (n.d.) 177–182. doi:10.1021/ci049714+.
- [63] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, M.K. Gilson, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities., *Nucleic Acids Res.* 35 (2007) D198–D201. doi:10.1093/nar/gkl999.
- [64] A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, et al., The ChEMBL bioactivity database: an update., *Nucleic Acids Res.* 42 (2014) D1083–1090. doi:10.1093/nar/gkt1031.
- [65] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, et al., DrugBank 4.0: shedding new light on drug metabolism., *Nucleic Acids Res.* 42 (2014) D1091–D1097. doi:10.1093/nar/gkt1068.
- [66] Drug and target protein structures in the PDB, (n.d.). <http://www.ebi.ac.uk/thornton-srv/databases/drugport/> (accessed September 18, 2014).

- [67] P. Romero, J. Wagg, M. Green, D. Kaiser, M. Krummenacker, P. Karp, Computational prediction of human metabolic pathways from the complete human genome, *Genome Biol.* 6 (2004) R2. doi:10.1186/gb-2004-6-1-r2.
- [68] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, et al., HMDB 3.0--The Human Metabolome Database in 2013., *Nucleic Acids Res.* 41 (2013) D801–D807. doi:10.1093/nar/gks1065.
- [69] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, Data, information, knowledge and principle: back to metabolism in KEGG., *Nucleic Acids Res.* 42 (2014) D199–205. doi:10.1093/nar/gkt1076.
- [70] MDDR, (n.d.).
<http://accelrys.com/products/databases/bioactivity/mddr.html> (accessed September 18, 2014).
- [71] J. Nickel, B.-O. Gohlke, J. Erehman, P. Banerjee, W.W. Rong, A. Goede, et al., SuperPred: update on drug classification and target prediction., *Nucleic Acids Res.* 12 (2014) W26–W31. doi:10.1093/nar/gku477.
- [72] L. Wang, C. Ma, P. Wipf, H. Liu, W. Su, X. Xie, TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database., *AAPS J.* 15 (2013) 395–406. doi:10.1208/s12248-012-9449-z.
- [73] J. Gong, C. Cai, X. Liu, X. Ku, H. Jiang, D. Gao, et al., ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method., *Bioinformatics.* 29 (2013) 1827–1829. doi:10.1093/bioinformatics/btt270.
- [74] S. Kim Kjørulff, L. Wich, J. Kringelum, U.P. Jacobsen, I. Kouskoumvekaki, K. Audouze, et al., ChemProt-2.0: visual navigation in

- a disease chemical biology database., *Nucleic Acids Res.* 41 (2013) D464–D469. doi:10.1093/nar/gks1166.
- [75] J.-C. Wang, P.-Y. Chu, C.-M. Chen, J.-H. Lin, idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach., *Nucleic Acids Res.* 40 (2012) W393–W399. doi:10.1093/nar/gks496.
- [76] A. Lagunin, A. Stepanchikova, D. Filimonov, V. Poroikov, PASS: prediction of activity spectra for biologically active substances., *Bioinformatics.* 16 (2000) 747–748.

Figures

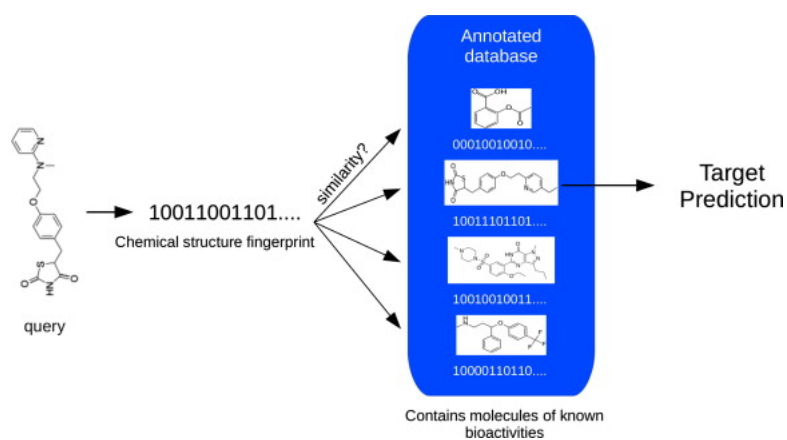


Figure 1. Chemical similarity through the comparison of fingerprints can be used to predict novel targets or functions of a query molecule.

Table 1. Databases and web resources useful for *in silico* target fishing.

| Database | URL | Description | Reference |
|---------------------------|---|---|-----------|
| BindingDB | http://www.bindingdb.org/bind/ | Database of measured binding affinities, focusing primarily on the interactions of proteins considered to be drug-targets with small, drug-like molecules | [63] |
| ChEMBL | http://www.ebi.ac.uk/chembl | Contains 2D structures, calculated properties, and abstracted bioactivities of drug-like small molecules | [64] |
| DrugBank | http://www.drugbank.ca/ | Contains information about drugs and drug targets | [65] |
| DrugPort | http://www.ebi.ac.uk/thornton-srv/databases/drugport/ | Provides an analysis of the structural information available in the PDB relating to drug molecules and their protein targets | [66] |
| HumanCyc | http://humancyc.org/ | Provides an encyclopedic reference and computer-queryable database of human metabolic pathways | [67] |
| Human Metabolome Database | http://www.hmdb.ca/ | Contains detailed information about small-molecule metabolites found in the human body | [68] |
| KEGG | http://www.genome.jp/kegg/ | Integrated database resource containing information about pathway maps, metabolites, small molecules, and drugs | [69] |
| MDL Drug Data Report | http://accelrys.com/products/databases/bioactivity/mddr.html | Contains over 150,000 biologically relevant compounds and well-defined derivatives | [70] |
| PubChem | http://pubchem.ncbi.nlm.nih.gov/ | Open repository for experimental data identifying the biological activities of small molecules | [59] |
| SuperTarget | http://bioinf-apache.charite.de/supertarget_v2/ | Extensive web resource for analyzing drug-target interactions | [60] |
| WOMBAT | http://www.sunsetmolecular.com/ | Contains chemical series from published literature | [61] |
| ZINC | https://zinc.docking.org | Free database of commercially available compounds for virtual screening. It provides subsets with actives for many ChEMBL targets. | [62] |

Table 2. Web servers that can be employed for *in silico* target fishing

| Server | Method | URL | Reference |
|------------------------------------|--|---|-----------|
| ChemMapper | 3D similarity (molecular shape and chemotype features) computation | http://lilab.ecust.edu.cn/chemmapper/ | [73] |
| ChemProt 2.0 | Fingerprint based | http://www.cbs.dtu.dk/services/ChemProt-2.0/ | [74] |
| DRAR-CPI | Docking based | http://cpi.bio-x.cn/drar/ | [41] |
| HitPick | Combines 2D fingerprints and a machine learning method based on a Laplacian-modified naive Bayesian model | http://mips.helmholtz-muenchen.de/proj/hitpick | [51] |
| idTarget | Divide-and-conquer docking approach | http://idtarget.rcas.sinica.edu.tw/ | [75] |
| Mantra 2.0 | Network theory and non-parametric statistics on gene expression data | http://mantra.tigem.it/ | [48] |
| PASS ONLINE | Fingerprint based Bayesian approach based on the knowledge base about structure-activity relationships for more than 260,000 compounds | http://www.pharmaexpert.ru/passonline/ | [76] |
| PharmMapper | Pharmacophore based | http://59.78.96.61/pharmmapper/index.php | [43] |
| Similarity ensemble approach (SEA) | Fingerprint based | http://sea.bkslab.org/ | [37] |
| SPiDER | Self-organizing map-based prediction | http://modlab-cadd.ethz.ch/software/spider/ | [38] |
| SuperPred | A combination of 2D, 3D, and fragment similarity values | http://prediction.charite.de/ | [71] |
| SwissTargetPrediction | A combination of 2D and 3D similarity values | http://www.swisstargetprediction.ch/ | [25] |
| TarFisDock | Docking based | http://www.dddc.ac.cn/tarfisdock/ | [40] |
| TargetHunter | Fingerprint based | http://www.cbligand.org/TargetHunter/ | [72] |

Anglerfish: a webserver for quantitative prediction of ligand bioactivity.

Adrià Cereto-Massagué^{1,2}, María José Ojeda¹, Aleix Gimeno¹, Sarah Tomás-Hernández¹, Raúl Beltrán-Debón¹, Cristina Valls¹, Miquel Mulero¹, Santiago Garcia-Vallve^{1,3}, Gerard Pujadas^{1,3}*

¹ Research group in Cheminformatics & Nutrition, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus Sescelades, 43007 Tarragona, Catalonia, Spain

² Group of Research on Omic Methodologies (GROM). Joint Unit Universitat Rovira i Virgili - EURECAT Technology Centre of Catalonia. Centre for Omic Sciences (COS), Unique Scientific and Technical Infrastructures (ICTS), Avinguda Universitat, 1, 43204 Reus, Catalonia, Spain

³ EURECAT, TECNIO, CEICS, Avinguda Universitat, 1, 43204 Reus, Catalonia, Spain

* Corresponding author: santi.garcia-vallve@urv.cat

Keywords

Computational Target Fishing, Reverse Screening, Drug Repositioning, Polypharmacology, Drug Repurposing, Drug Reprofilng, Target Prediction, Activity Prediction.

Abstract

Anglerfish is a new freely available web server that does fast target prediction for small molecules. It does so by combining several different molecular fingerprints (which can be chosen by the user) and by leveraging ChEMBL activity data to predict potential new targets for the query molecule and quantifying its potential bioactivity value towards those targets (measured as a standardized pX value). The method has been validated using data from different ChEMBL versions, by being able to reliably predict activities from a newer ChEMBL version by using exclusively data from an older version and a combination of three fingerprints (*i.e.*, RDKit Fingerprint and FP3 and MACCS166 from OpenBabel). Anglerfish can be used for free at <http://anglerfish.urv.cat>.

Introduction

The fast growth of freely available bioactivity data resources like PubChem¹⁻³, BindingDB^{4,5} or ChEMBL⁶ during the last decade has enabled the development of approaches that leverage all these data in order to predict bioactivities.

One of such approaches is Target Fishing, also known as *reverse pharmacognosy*, *reverse screening*, *polypharmacology prediction*, *drug repurposing*, *drug reprofiling* or *target prediction*⁷, which is the opposite approach to that of Virtual Screening^{8,9}: in Virtual Screening, bioactive compounds for a given target are sought; in Target Fishing the starting point is the small molecule, for which a target is sought. It is thus a very appealing

approach for drug repurposing (with the associated benefits when compared to developing a new drug¹⁰) and in order to study known or prospective drugs' potential adverse effects or polypharmacology (which is likely, given that, on average, each known drug has 6 different molecular targets¹¹). Current Target Fishing methods can be broadly classified in two categories, pretty much like Virtual Screening: target-centric methods, which often rely on machine learning models built on top of known activity data for a set of targets, with some that take the structural models of the targets into account or even dock the query compounds to a wide range of targets; and ligand-centric methods, which rely on some similarity metric between the query molecule(s) and some known bioactive molecule⁷. Ligand-centric Target Fishing methods have the advantage of requiring much less known information about the targets, being that a single known active compound can be enough to identify a new putative target for a query molecule, while for target-centric methods either the structural model of the target or a sizeable amount of actives and inactives (in order to build prediction models) need to be known⁷.

Among the molecular similarity metrics used for ligand-centric approaches both in Target Fishing and in Virtual Screening, molecular fingerprint Tanimoto similarity index¹² is a common choice that enables the easy comparison of compounds by taking into account different sets of their features (depending on the fingerprinting algorithm chosen¹³). There are some freely accessible Target Fishing that implement molecular fingerprint similarity search⁷, with some even enabling the user to choose which fingerprint to use¹⁴. The combination, or data fusion, of the similarities from different molecular fingerprints has proven to increase Virtual Screening performance¹⁵, but this has not yet been leveraged by any freely available Target Fishing tool or service. Thus, we provide a novel Target Fishing approach leveraging several

simultaneous molecular fingerprint similarities and ChEMBL⁶ activity data fusion, with a fast and freely accessible online implementation: Anglerfish.

Implementation and Methods

Molecule Database

The molecule database used in Anglerfish was derived from a subset of ChEMBL version 22. This subset consists of all compounds with Pubchem Bioassays of type B (data measuring binding of compound to a molecular target) whose target was a protein or protein complex, its activity could be expressed in nM (as “standard unit” in ChEMBL) and whose standard type was either of AC50, EC50, ED50, IC50, ID50, Ka, Kb, Kd, Ke, Ki, LC50, LD50, Potency or XC50. All these different activity measures were normalized into a pX value equivalent to ChEMBL’s pChEMBL⁶, calculated as $9 - \log(\textit{standard_value})$, where *standard_value* is the representation of the activity in nM. When more than one activity was found for a single compound-target pair, all values were averaged.

Molecular fingerprints of each of the supported types (9 at present) are calculated for each molecule in the database and stored for similarity searches.

Algorithm

Provided a query molecule and a given set of molecular fingerprints (which are discussed below), each fingerprint is calculated for the query and then a similarity search against the Anglerfish database is performed. The results of these similarity searches are then combined, and for each active hit an average

similarity measure between the used fingerprints is calculated. Using these similarities and the ChEMBL activities of these molecules, a pX activity value for the query molecule is predicted for each target.

The available fingerprints are:

- Key-based fingerprints:
 - MACCS166¹⁶ (OpenBabel and RDKit implementations)
 - OpenBabel FP3¹⁷
 - OpenBabel FP4¹⁷
- Topological (path-based) fingerprints:
 - RDKit topological fingerprint (daylight-like)¹⁸
 - RDKit Torsion¹⁹
 - RDKit Atom Pairs²⁰
 - OpenBabel FP2 (daylight-like)¹⁷
- Circular fingerprints
 - RDKit Morgan (ECFP)²¹

Similarity searches are performed through the ChempFP²² library. Cinfony²³, pybel²⁴ and Indigo²⁵ are used internally too.

Interface

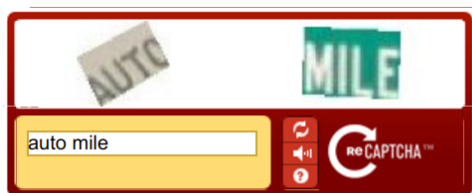
Anglerfish provides a simple and an easy to use web-based front-end. On the landing site (see Figure 1), the user can provide a query molecule by 3 possible means: providing a SDF file containing a single molecule, writing its SMILES, or writing its InChI²⁶ code. Following this, the user can also select which molecular fingerprints to use for the similarity search.

Figure 8. Anglerfish landing page, from where a new search can be launched

Target Fishing

You can draw molecules using the [PubChem Sketcher](#) and then pasting their SMILES or InChI here.

Captcha:



SDF File:

No file chosen

SMILES or InChi string:


COC1=CC=C(C=C1)C(Cl)=C(C1=CC=C(OC)C=C1)C1=C
C=C(OC)C=C1 Chlorotrianisene example

Query
molecule
input

- MACCS166 (RDKit)
- RDKit Fingerprint
- Morgan (RDKit)
- Torsion (RDKit)
- AtomPair (RDKit)
- FP2 (OpenBabel)
- FP3 (OpenBabel)
- FP4 (OpenBabel)
- MACCS166 (OpenBabel)

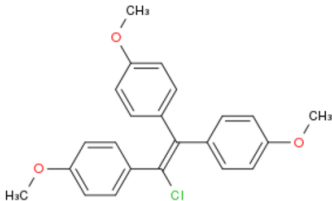
Molecular
fingerprints to use

After submitting this form, a confirmation screen (see Figure 2) will be shown to make sure everything is correct. Once confirmed, the activity search will start, and the user will be redirected to a waiting page showing the status of the search. The user will be provided with a unique URL pointing to that page, which will show the search results once they are available (usually in a couple of minutes). This URL can be saved for future reference to the search results.



Anglerfish

» Search » Help



Molecule: Chlorotrianisene example
Fingerprints:
RDKit-Fingerprint , OpenBabel-FP3 , OpenBabel-MACCS

SMILES:
COC1C=CC(=CC=1)C(Cl)=C(C1C=CC(=CC=1)OC)C1C=CC(=CC=1)OC

Confirmation

Proceed Back

Figure 9. The confirmation page lets the user review the chosen parameters before continuing

Once the results are ready, they will automatically load and they will be shown as a table (see Figure 3) with the following fields: target name, predicted pX, highest average similarity, lowest average similarity and hits (number of similar active compounds for that target). This table is interactive and can be sorted by any of the columns and it also allows for complex search criteria in its search box. By default it is sorted by descending average similarity and then by the predicted pX in descending order, meaning the targets for which the predicted activity is most likely will show at the top of the results. The results table can be downloaded as a CSV spreadsheet.

Clicking on the “Hits” column leads to a detailed view of the similarities between the query and each active compound for that target (see Figure 4).

Perform new search

» Search » Help

Results

To help page

Result URL
<http://anglerfish.urv.cat/anglerfish/results?id=0b501ba3-4u7a0f057yd0f-743a0d047c-75f8u3cu7y>

Fingerprints used: RDKit-Fingerprint, OpenBabel-FP3, OpenBabel-MACCS

Repeat

Search parameters

Query molecule

Filter results

Search Clear

2718 records found

| Target Name | Predicted pX | Max Average Similarity | Min Average Similarity | Hits |
|---|--------------|------------------------|------------------------|-----------------------|
| Ferritin light chain | 4.80 | 1.00 | 0.09 | 15897 |
| Androgen Receptor | 4.37 | 0.84 | 0.09 | 1098 |
| Quinone reductase 2 | 4.62 | 0.74 | 0.09 | 163 |
| Cyclooxygenase-1 | 5.15 | 0.73 | 0.09 | 1977 |
| Aryl hydrocarbon receptor | 6.32 | 0.72 | 0.09 | 86 |
| Estrogen receptor alpha | 6.31 | 0.72 | 0.09 | 1584 |
| Protein kinase C alpha | 3.96 | 0.72 | 0.09 | 890 |
| Nuclear factor NF-kappa-B p65 subunit | 4.59 | 0.71 | 0.09 | 101 |
| Sortase | 3.25 | 0.71 | 0.09 | 49 |
| 3-beta-hydroxysteroid-delta(8),delta(7)-isomerase | 6.39 | 0.71 | 0.09 | 82 |
| C-8 sterol isomerase | 6.17 | 0.71 | 0.09 | 17 |
| Sigma opioid receptor | 5.89 | 0.71 | 0.09 | 857 |
| Serotonin transporter | 5.08 | 0.71 | 0.09 | 2932 |
| Serotonin 6 (5-HT6) receptor | 5.03 | 0.71 | 0.09 | 872 |
| HERG | 4.78 | 0.71 | 0.09 | 3039 |
| Dopamine transporter | 4.74 | 0.71 | 0.09 | 1962 |
| Alpha-2c adrenergic receptor | 4.73 | 0.71 | 0.09 | 276 |
| Muscarinic acetylcholine receptor M4 | 4.71 | 0.71 | 0.09 | 211 |
| Dopamine D3 receptor | 4.62 | 0.71 | 0.09 | 1454 |
| Alpha-2a adrenergic receptor | 4.58 | 0.71 | 0.09 | 414 |

To detailed result table

Result table

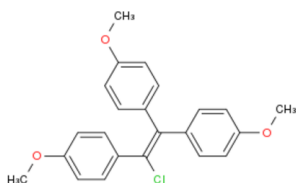
Export:CSV

Download whole result table

1,2,3,4,5, >>>

Further result pages

Figure 10. View of the results page



results.target_name = "f"

15997 records found

| Target Name | Standard Type | pX | Active | Average Similarity | Rdkit Tanimoto | Fingerprint Tanimoto | Openbabel Tanimoto | Openbabel Tanimoto | Maccs Tanimoto |
|---|---------------|------|-------------------------------|--------------------|----------------|----------------------|--------------------|--------------------|----------------|
| Ferritin <small>light chain</small> | Potency | 4.80 | CHEMBL1200761 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | |
| Ferritin <small>light chain</small> | Potency | 4.30 | CHEMBL1598215 | 0.87 | 0.75 | 1.00 | | 0.87 | |
| Ferritin <small>light chain</small> | Potency | 4.75 | CHEMBL1432495 | 0.65 | 0.39 | 1.00 | | 0.55 | |
| Ferritin <small>light chain</small> | Potency | 4.30 | CHEMBL69003 | 0.63 | 0.39 | 1.00 | | 0.52 | |
| Ferritin <small>light chain</small> | Potency | 4.30 | CHEMBL1368968 | 0.62 | 0.35 | 1.00 | | 0.52 | |
| Ferritin <small>light chain</small> | Potency | 4.45 | CHEMBL1386064 | 0.62 | 0.35 | 1.00 | | 0.52 | |
| Ferritin <small>light chain</small> | Potency | 4.38 | CHEMBL1395591 | 0.62 | 0.35 | 1.00 | | 0.50 | |
| Ferritin <small>light chain</small> | Potency | 4.35 | CHEMBL1433715 | 0.62 | 0.35 | 1.00 | | 0.50 | |
| Ferritin <small>light chain</small> | Potency | 4.30 | CHEMBL1481093 | 0.62 | 0.35 | 1.00 | | 0.50 | |
| Ferritin <small>light chain</small> | Potency | 4.95 | CHEMBL1344868 | 0.62 | 0.35 | 1.00 | | 0.50 | |
| Ferritin <small>light chain</small> | Potency | 4.55 | CHEMBL1318402 | 0.62 | 0.35 | 1.00 | | 0.50 | |
| Ferritin <small>light chain</small> | Potency | 4.80 | CHEMBL454843 | 0.61 | 0.36 | 1.00 | | 0.49 | |
| Ferritin <small>light chain</small> | Potency | 4.30 | CHEMBL1341404 | 0.61 | 0.34 | 1.00 | | 0.50 | |
| Ferritin <small>light chain</small> | Potency | 4.40 | CHEMBL1320529 | 0.61 | 0.35 | 1.00 | | 0.49 | |
| Ferritin <small>light chain</small> | Potency | 4.30 | CHEMBL1329576 | 0.61 | 0.35 | 1.00 | | 0.49 | |
| Ferritin <small>light chain</small> | Potency | 4.45 | CHEMBL1306311 | 0.61 | 0.36 | 1.00 | | 0.46 | |

Figure 11. Detailed results view, showing each activity match and their individual similarities

Validation

For the validation of the method, a new database was built, using the same procedure, with the previous version of ChEMBL (version 21). A set of 217 molecules were selected from the database built from the latest ChEMBL, such as:

- They were different enough between them according to all tested fingerprints (Tanimoto < 0.7)
- They were not present in ChEMBL 21
- Their targets were present in ChEMBL 21, and they had active compounds that could be included in the Anglerfish database

This way, the predicted activities of these compounds against the ChEMBL21-based database could be compared against their known actual activities from ChEMBL22.

For the predicted activity, we tested 3 different formulas:

1. The first one was the average activity weighted with the average similarities:

$$\frac{\sum pX_i \times \text{avg_tanimoto}_i}{\sum \text{avg_tanimoto}_i}$$

Where *avg_tanimoto* is the average of the Tanimoto similarity between the query and the active molecule for all the fingerprints chosen.

This formula comes from the assumption that the average Tanimoto similarity to an active will show how much of that activity can we expect on the query

molecule. With this formula, a molecule that shows the highest similarity to known actives will have the average of the activities of those actives as its predicted activity. Lower similarities will lower the predicted activity value, but will also contribute less to it. However, it can still be the case that a lot of low-similarity and low-activity known actives can affect negatively the predicted activity. There is also the added problem that whenever there are more than one known active, the predicted activity cannot be as high as the highest known activity, even if it is the only one from a very similar known active. This formula is also sensitive to the number of known actives: the more known actives a target has, the more the predicted activity tends to a lower value. For this reason, we also tried the next formula, in an attempt to avoid these problems.

2. The highest product of activity and average similarity:

$$\max(pX_i \times \text{avg_tanimoto}_i)$$

In this case, the predicted activity is the highest product of the similarity with the query and the activity value (pX) of a known active. This takes into account both the activity and the similarity to the query, but is insensitive to noise caused by too many known actives or too many dissimilar actives. This formula however has the problem of being potentially too optimistic for the predicted activity, since a very potent known active could hide many actually more similar but not so active results. This, again, prompted us to develop yet another formula to predict the activity.

3. The weighted activity scaled by the average similarity that produces the highest product of activity and average similarity:

$$\frac{\sum pX_i \times \text{avg_tanimoto}_i}{\sum \text{avg_tanimoto}_i} \times \text{avg_tanimoto}_j;$$

$$\text{for avg_tanimoto}_j \times pX_j = \max(pX_i \times \text{avg_tanimoto}_i)$$

In this last formula, we tried to combine the prior two methods, basically by multiplying the result of the first formula by the Tanimoto similarity that yield the highest product with its activity (the second formula).

The validation was tested for all fingerprints and all possible fingerprint combinations of up to 3 different fingerprints, yielding a total of 129 different combinations. Performance of the predictions was assessed with both Kendall's tau-b coefficient and the % of predictions within 1 pX of the known value.

Results and Discussion

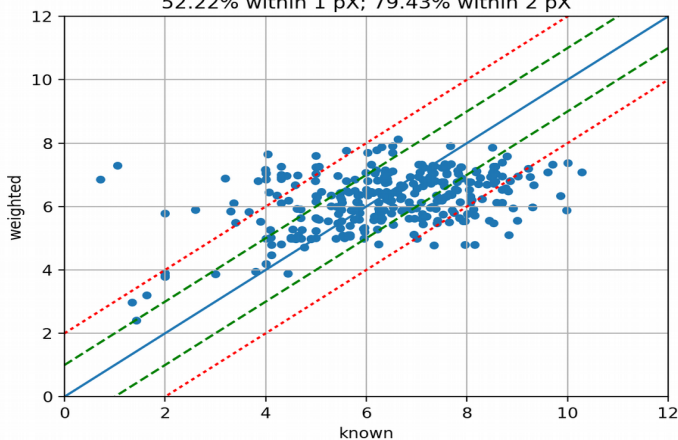
The predicted activities were plotted against their known counterparts in scatter plots, like those on Figure 5. The differences between the predicted values and the known values were plotted as Gaussian distributions, such as those in Figure 6. The plots for every fingerprint combination and formula can be found in the supplementary data.

Comparing the validation set predicted activities to their known activities, we can see in figure 5 that both the first and third formulas tend to predict activities within a certain range regardless of the known activity, yielding a low correlation between them. The second formula, however, shows higher correlation between predicted and known activities across different fingerprint combinations. It is with the second formula too that we find the 2 fingerprint combinations with the μ closest to 0 (≈ -0.04), that is, those whose average difference between predicted and known values are closer to 0. These 2 combinations are essentially the same: the RDKit's daylight-like fingerprint, OpenBabel's key-based FP3 and either of the MACCS166 implementations (OpenBabel or RDKit). With either of these combinations, 50% of the predicted activities are within 1 pX unit of their known values, which is increased to an 80% within 2 pX units of the known activities, as can be seen in Figures 5 and 6.

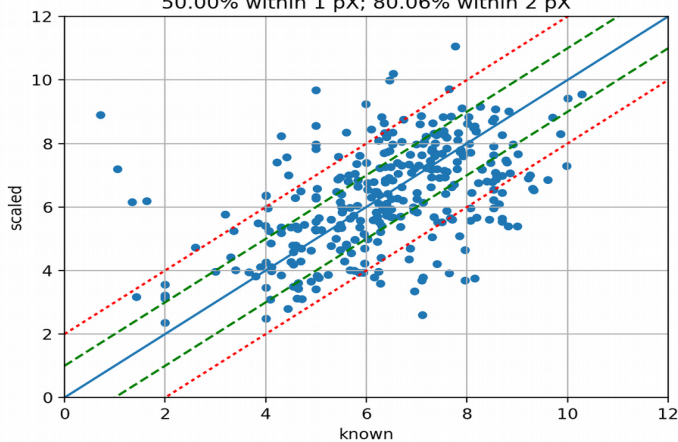
Based on these results, Anglerfish uses the second formula for activity prediction, and those fingerprints are enabled by default.

Figure 12. Predicted activity pX vs known activity for the combination of RDKit, FP3, and (OpenBabel) MACCS fingerprints, from top to bottom using the formulas 1 (“weighted”), 2 (“scaled”) and 3 (“expected”) for the predicted pX value. The green and red lines delimit the intervals within 1 and 2 pX of the known value, respectively. The blue line intersects the exact matches.

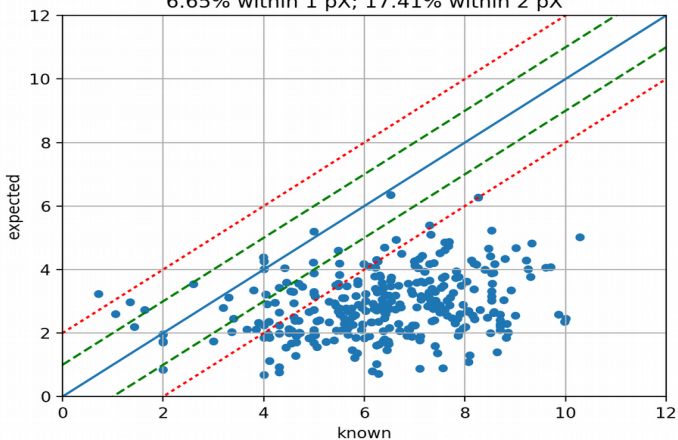
RDKit-Fingerprint+OpenBabel-FP3+OpenBabel-MACCS: Kendall $\tau = 0.26$
52.22% within 1 pX; 79.43% within 2 pX



RDKit-Fingerprint+OpenBabel-FP3+OpenBabel-MACCS: Kendall $\tau = 0.36$
50.00% within 1 pX; 80.06% within 2 pX



RDKit-Fingerprint+OpenBabel-FP3+OpenBabel-MACCS: Kendall $\tau = 0.18$
6.65% within 1 pX; 17.41% within 2 pX



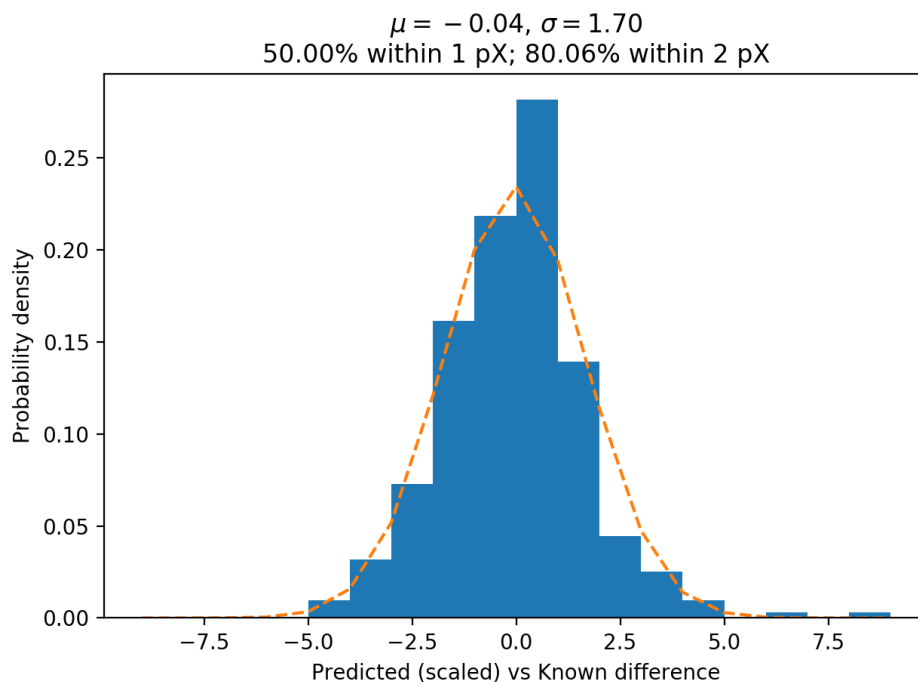


Figure 13. Distribution of the deviation of the predicted activities from their known using the second formula (“scaled”) values for the same fingerprints as in figure 5. Plots for the rest of formulas and fingerprint combinations can be found in the supplementary data.

An example target search was done borrowing chlorotrianisene from SwissTargetPrediction²⁷ as the example molecule, since it has some described activities not present in ChEMBL⁶ for some target which are present in ChEMBL: Cyclooxygenase-1 (COX-1)²⁸ and Estrogen Receptor alpha (ER)²⁹. As we can see from the results in Figure 3, besides the match against itself that gives the Ferritin light chain activity, we can find among the first predicted targets both COX-1 and ER. The second listed predicted target, Androgen

Receptor, has described activity within ChEMBL for this molecule (from PubChem bioassay³⁰), though as a functional assay (type “F”) and without activity values usable by Anglerfish. This leaves at the very least 3 true positives (without counting the self-match) among the first results.

Conclusions

Most target fishing software solutions and on-line tools make use of 2D similarity methods⁷, specially molecular fingerprints. While others combine a single 2D fingerprint with other different approaches for target fishing, Anglerfish is the only one to combine several different 2D molecular fingerprints, which provides it with very fast similarity searching and thus fast output results (usually available after just a couple of minutes). The novelty in Anglerfish is also in not just predicting potential new targets for a given molecule, but also in estimating its activity value; this can help prioritize targets with a potentially higher activity or avoid those with a very low predicted activity.

Supporting Information.

A zip archive is provided containing the detailed results from the validation process, for each of the activity prediction methods and the set of molecules used. Prediction scatter plots and deviation distribution plots are also provided.

Author Information

Corresponding Author

*Santiago Garcia-Vallve. Research group in Cheminformatics & Nutrition, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Campus Sescelades, 43007 Tarragona, Catalonia, Spain; e-mail: santi.garcia-vallve@urv.cat

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

This study was supported by research grants 2014PFR-URV-B2-67 and 2015PFR-URV-B2-67 from our University. AG's contract is supported by grant 2015FI_B00655 from the Catalonia Government.

References

- (1) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. In *Annual Reports in Computational Chemistry*; American Chemical Society: Washington, DC, 2008; Vol. 4, pp 217–241.
- (2) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. *Nucleic Acids Res.* **2017**, *45* (D1), D955–D963.
- (3) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213.
- (4) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. *Nucleic Acids Res.* **2007**, *35* (Database issue), D198–D201.
- (5) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. *Nucleic Acids Res.* **2016**, *44* (D1), D1045–D1053.
- (6) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. *Nucleic Acids Res.* **2014**, *42* (Database issue), D1083-1090.

- (7) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. *Methods* **2015**, *71*, 98–103.
- (8) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. *J. Chem. Inf. Model.* **2012**, *52* (4), 867–881.
- (9) Rester, U. *Curr. Opin. Drug Discov. Devel.* **2008**, *11* (4), 559–568.
- (10) Bisson, W. H. *Curr. Top. Med. Chem.* **2012**, *12* (17), 1883–1888.
- (11) Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. *Mol. Biosyst.* **2009**, *5* (9), 1051.
- (12) Bajusz, D.; Rácz, A.; Héberger, K. *J. Cheminform.* **2015**, *7* (1), 20.
- (13) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. *Methods* **2015**, *71*, 58–63.
- (14) Awale, M.; Reymond, J.-L. *Nucleic Acids Res.* **2014**, gku379-.
- (15) Salim, N.; Holliday, J.; Willett, P. *J. Chem. Inf. Comput. Sci.* **2002**, *43* (2), 435–442.
- (16) Accelrys. San Diego, CA.
- (17) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform.* **2011**, *3*, 33.
- (18) Landrum, G. RDKit: Open-source cheminformatics
<http://www.rdkit.org>.

(19) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. *J. Chem. Inf. Model.* **1987**, *27* (2), 82–85.

(20) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Model.* **1985**, *25* (2), 64–73.

(21) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

(22) Dalke, A. *J. Cheminform.* **2013**, *5* (Suppl 1), P36.

(23) O’Boyle, N. M.; Hutchison, G. R. *Chem. Cent. J.* **2008**, *2* (1), 24.

(24) O’Boyle, N. M.; Morley, C.; Hutchison, G. R. *Chem. Cent. J.* **2008**, *2* (1), 5.

(25) Indigo - EPAM Life Sciences
<http://lifescience.opensource.epam.com/indigo/>.

(26) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *J. Cheminform.* **2015**, *7* (1), 23.

(27) Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. *Nucleic Acids Res.* **2014**, *42* (W1), W32–W38.

(28) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. *Nature* **2012**.

(29) Kupfer, D.; Bulger, W. H. *FEBS Lett.* **1990**, *261* (1), 59–62.

(30) National Center for Biotechnology Information. PubChem BioAssay Database. AID 743053 - qHTS assay to identify small molecule agonists of the

androgen receptor (AR) signaling pathway: Summary - PubChem BioAssay

Summary <https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=743053>

(accessed May 22, 2017).

Conclusions

The literature reviews that are part of this thesis have been extensively cited in all kinds of papers in the field, both for molecular fingerprint similarity search in virtual screening¹⁻³⁵ and for *in silico* target fishing^{15,36-56}. But most interesting is the fact that the published tools developed as part of this thesis have been proven useful by the cheminformatics community:

The first objective of this thesis (“*To create a freely available tool with a graphical interface to facilitate the validation of virtual screening approaches by making decoy molecule library building easier and more accessible.*”) was tackled by developing DecoyFinder, which has been widely downloaded and used throughout the world for the validation of virtual screenings⁵⁷⁻⁸⁹, and it has also been an inspiration for others to implement their own in-house algorithms⁹⁰⁻⁹⁷, improve available resources⁹⁸ and even its code has been used as the foundation for a decoy library building web service⁹⁹. It has also been featured in a number of reviews of the field¹⁰⁰⁻¹⁰⁷ and books^{89,108}. Of note is its use in furthering research against diseases such as malaria^{61,65,78,82}, cancer^{58,67,68,80,81,88}, visceral leishmaniasis^{60,63,69}, HIV⁸³, hepatitis C⁶², Alzheimer’s^{85,86}, Parkinson’s^{57,86}, and diabetes^{74,77,79,89}, among others and other health-related research, such as on novel antibacterial agents^{64,66,73}.

The results of from the *Molecular weight-based decoys: a simple decoy set finding alternative for fingerprint similarity approaches* manuscript have been integrated into DecoyFinder. However, since it has not yet been published, its future impact cannot be assessed, but it has the potential to further spread the

usage of DecoyFinder by turning it more suitable for ligand-based virtual screening approaches.

The development of VHELIBS was the answer to the second objective (“*To develop a publicly available target fishing service with quantitative bioactivity prediction.*”), which was dealt with with the collaboration of the creators of the PDB_REDO. It has also been widely downloaded, and although this has not translated yet in such a wide range of publications as DecoyFinder, besides being directly used for the selection of appropriate structure models^{89,109}, it has also been used to validate the development of a new virtual screening web service¹¹⁰ and to assess the overall reliability in general of the available protein-ligand crystallography models¹¹¹. It has also been positively featured into many reviews of the field^{107,108,112–122}. In addition to this, VHELIBS has been used to teach undergraduate students about the quality features of crystallographic Protein Data Bank structures and how to assess them.

Both tools have clearly filled a need in their respective niches, having a significant impact too, and thus, can be considered a success.

The last objective of this thesis (“*To develop a publicly available target fishing service with quantitative bioactivity prediction*”) led to the development of Anglerfish, which was just recently made publicly available and does not have yet a published paper, which means that at the moment of this writing it still has not had the chance to have a meaningful impact in the field. However, given its characteristics, hopefully it will be seen as a valuable resource in the cheminformatics ecosystem in the near future.

References

- (1) Perricone, U.; Wieder, M.; Seidel, T.; Langer, T.; Padova, A.; Almerico, A. M.; Tutone, M. *ChemMedChem* **2017**.
- (2) Himmat, M.; Salim, N.; Al-Dabbagh, M.; Saeed, F. *Molecules* **2016**.
- (3) VanPatten, S.; Sun, S.; He, M.; Cheng, K. *J. Med.* **2016**.
- (4) Muegge, I.; Mukherjee, P. *Expert Opin. Drug Discov.* **2016**, *11* (2), 137–148.
- (5) Duesbury, E. Applications and Variations of the Maximum Common Subgraph for the Determination of Chemical Similarity, 2015.
- (6) Skoda, P.; Hoksza, D. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; IEEE, 2016; pp 1220–1227.
- (7) Zhang, L.; Ai, H.; Chen, W.; Yin, Z.; Hu, H.; Zhu, J.; Zhao, J.; Zhao, Q.; Liu, H. *Sci. Rep.* **2017**, *7* (1), 2118.
- (8) Dong, J.; Cao, D.-S.; Miao, H.-Y.; Liu, S.; Deng, B.-C.; Yun, Y.-H.; Wang, N.-N.; Lu, A.-P.; Zeng, W.-B.; Chen, A. F. *J. Cheminform.* **2015**, *7* (1), 60.
- (9) O’Boyle, N.; Sayle, R. *J. Cheminform.* **2016**.
- (10) Borrel, A. Development of Computational Methods to Predict Pocket Druggability and Profile Ligands using Structural, 2016.
- (11) Skinnider, M. A.; Johnston, C. W.; Edgar, R. E.; Dejong, C. A.; Merwin, N. J.; Rees, P. N.; Magarvey, N. A. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (42), E6343–E6351.
- (12) Garcia-Castro, M.; Zimmermann, S. *Angewandte* **2016**.
- (13) Zulkefli, Z. B.; Malim, N. H. A. H.; Al-Laila, M. H. In *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*; IEEE, 2016; pp 282–286.

- (14) Wang, L.; Forni, F.; Ortega, R.; Liu, Z.; Su, H. *IEEE Trans. Automat. Contr.* **2016**, *PP*, 1–1.
- (15) Minkiewicz, P.; Darewicz, M.; Iwaniak, A.; Bucholska, J.; Starowicz, P.; Czyrko, E. *Int. J. Mol. Sci.* **2016**, *17* (12), 2039.
- (16) Rogo, M. Modeling and synthesis of antiplasmodial benzoxazines from natural products of Kenya, 2016.
- (17) Sarupinda, T. I. Modelling cellular permeability via carrier mediated transport, 2015.
- (18) Neves, B. J.; Muratov, E.; Machado, R. B.; Andrade, C. H.; Cravo, P. V. L. *Expert Opin. Drug Discov.* **2016**, *11* (6), 557–567.
- (19) Vass, M.; Kooistra, A. J.; Ritschel, T.; Leurs, R.; de Esch, I. J.; de Graaf, C. *Curr. Opin. Pharmacol.* **2016**, *30*, 59–68.
- (20) B, W. K.; Pasupa, K. *Neural Information Processing*; Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2016; Vol. 9947.
- (21) Koes, D. *Comput. Drug Discov.* **2016**.
- (22) Warszycki, D.; Śmieja, M.; Kafel, R. *Mol. Divers.* **2017**.
- (23) Zhou, B.; Sun, Q.; Kong, D.-X. *Oncotarget* **2016**, *7* (22).
- (24) Janet, J. P.; Kulik, H. J. *Chem. Sci.* **2017**, 1–50.
- (25) Lee, A. A.; Brenner, M. P.; Colwell, L. J. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (48), 13564–13569.
- (26) Koutsoukas, A.; St. Amand, J.; Mishra, M.; Huan, J. *Front. Environ. Sci.* **2016**, *4*.
- (27) Kinjo, A. R.; Bekker, G.-J.; Suzuki, H.; Tsuchiya, Y.; Kawabata, T.; Ikegawa, Y.; Nakamura, H. *Nucleic Acids Res.* **2017**, *45* (D1), D282–D288.

- (28) Garcia-Castro, M.; Zimmermann, S.; Sankar, M. G.; Kumar, K. *Angew. Chemie Int. Ed.* **2016**, *55* (27), 7586–7605.
- (29) Daina, A.; Michielin, O.; Zoete, V. *Sci. Rep.* **2017**, *7*, 42717.
- (30) Zoete, V.; Daina, A.; Bovigny, C.; Michielin, O. *J. Chem. Inf. Model.* **2016**, *56* (8), 1399–1404.
- (31) Vilar, S.; Hripesak, G. *Brief. Bioinform.* **2016**, bbw048.
- (32) Rafati-Afshar, A. Unified processing framework of high-dimensional and overly imbalanced chemical datasets for virtual screening., 2017.
- (33) Lima, A. N.; Philot, E. A.; Trossini, G. H. G.; Scott, L. P. B.; Maltarollo, V. G.; Honorio, K. M. *Expert Opin. Drug Discov.* **2016**, *11* (3), 225–239.
- (34) Hoksza, D.; Skoda, P. In *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*; IEEE, 2016; pp 1–6.
- (35) Bajusz, D.; Rácz, A.; Héberger, K. *J. Cheminform.* **2015**, *7* (1), 20.
- (36) Wong, V. K.-W.; Law, B. Y.-K.; Yao, X.-J.; Chen, X.; Xu, S. W.; Liu, L.; Leung, E. L.-H. *Pharmacol. Res.* **2016**, *111*, 546–555.
- (37) Akhtar, F.; Sharif, H.; Mallick, M.; Zahoor, F. *Polish Pharm. Soc.* **2017**, *74* (2), 321–329.
- (38) Wu, X.; Chen, X.; Dan, J.; Cao, Y.; Gao, S.; Guo, Z.; Zerbe, P.; Chai, Y.; Diao, Y.; Zhang, L. *Sci. Rep.* **2016**, *6*, 25491.
- (39) Lo, Y.-C. Chemical Dissection of the Cell Cycle for Anticancer Drug Discovery and Target Identification, University of California, 2016.
- (40) Sol, C.; Rubio-Carrasco, K.; Díaz-Juárez, A.; Soledad; González-Covarrubias, V.; Fuentes-Noriega, I. *J. Mex. Chem. Soc.* **2014**, *58* (3), 287–302.
- (41) Qiu, Z.-C.; Dong, X.-L.; Dai, Y.; Xiao, G.-K.; Wang, X.-L.; Wong, K.-C.; Wong, M.-S.; Yao, X.-S. *Int. J. Mol. Sci.* **2016**, *17* (12), 2116.

- (42) Lauria, A.; Bonsignore, R.; Bartolotta, R.; Perricone, U.; Martorana, A.; Gentile, C. *Curr. Pharm. Des.* **2016**, *22* (21), 3073–3081.
- (43) Medina-Franco, J. L. *Epi-Informatics: discovery and development of small molecule epigenetic drugs and probes.*
- (44) Shmelkov, E.; Grigoryan, A.; Swetnam, J.; Xin, J.; Tivon, D.; Shmelkov, S. V.; Cardozo, T. *Front. Physiol.* **2015**, *6* (DEC).
- (45) Mori, M.; Cau, Y.; Vignaroli, G.; Laurenzana, I. *ACS Chem.* **2015**.
- (46) Peón, A.; Dang, C. C.; Ballester, P. J. *Front. Chem.* **2016**, *4*, 15.
- (47) Ye, X.-Y.; Ling, Q.-Z.; Chen, S.-J. *Evid. Based. Complement. Alternat. Med.* **2015**, *2015*, 983951.
- (48) Lo, Y.; Senese, S.; Li, C.; Hu, Q.; Huang, Y. *PLoS Comput* **2015**.
- (49) Huang, T.; Mi, H.; Lin, C.; Zhao, L.; Zhong, L. L. D.; Liu, F.; Zhang, G.; Lu, A.; Bian, Z. *BMC Bioinformatics* **2017**, *18* (1), 165.
- (50) Kapoor, S.; Waldmann, H.; Ziegler, S. *Bioorg. Med. Chem.* **2016**, *24* (15), 3232–3245.
- (51) Méndez-Lucio, O.; Naveja, J. J.; Vite-Caritino, H.; Prieto-Martínez, F. D.; Medina-Franco, J. L. *J. Mex. Chem. Soc.* **2016**, *60* (3), 168–181.
- (52) Silveira-Dorta, G.; Sousa, I. J.; Fernandes, M. X.; Martín, V. S.; Padrón, J. M. *Eur. J. Med. Chem.* **2015**, *96*, 308–317.
- (53) Chen, S.; Cui, M. *Molecules* **2017**.
- (54) Cruz-Monteagudo, M.; Schürer, S.; Tejera, E.; Pérez-Castillo, Y.; Medina-Franco, J. L.; Sánchez-Rodríguez, A.; Borges, F. *Drug Discov. Today* **2017**.
- (55) Awale, M.; Reymond, J.-L. *J. Cheminform.* **2017**, *9* (1), 11.
- (56) Luo, Q.; Zhao, L.; Hu, J.; Jin, H.; Liu, Z.; Zhang, L. *PLoS One* **2017**, *12* (2), e0171433.
- (57) Bhayye, S. S.; Roy, K.; Saha, A. *Med. Chem. Res.* **2014**, *23* (8), 3705–3713.

- (58) Xu, Y.; Yue, L.; Wang, Y.; Xing, J.; Chen, Z.; Shi, Z.; Liu, R.; Liu, Y. C.; Luo, X.; Jiang, H.; Chen, K.; Luo, C.; Zheng, M. *J. Chem. Inf. Model.* **2016**, *56* (9), 1847–1855.
- (59) Astolfi, A.; Iraci, N.; Sabatini, S.; Barreca, M.; Cecchetti, V. *Molecules* **2015**, *20* (9), 15842–15861.
- (60) Pandey, R. K.; Verma, P.; Sharma, D.; Bhatt, T. K.; Sundar, S.; Prajapati, V. K. *Biomed. Pharmacother.* **2016**, *83*, 141–152.
- (61) Kaalia, R.; Srinivasan, A.; Kumar, A.; Ghosh, I. *Mach. Learn.* **2016**, *103* (3), 309–341.
- (62) Wei, Y.; Li, J.; Qing, J.; Huang, M.; Wu, M.; Gao, F.; Li, D.; Hong, Z.; Kong, L.; Huang, W.; Lin, J. *PLoS One* **2016**, *11* (2), e0148181.
- (63) Pandey, R. K.; Kumbhar, B. V.; Sundar, S.; Kunwar, A.; Prajapati, V. K. *J. Recept. Signal Transduct.* **2017**, *37* (1), 60–70.
- (64) Perdih, A.; Hrast, M.; Pureber, K.; Barreteau, H.; Grdadolnik, S. G.; Kocjan, D.; Gobec, S.; Solmajer, T.; Wolber, G. *J. Comput. Aided. Mol. Des.* **2015**, *29* (6), 541–560.
- (65) Shah, P.; Tiwari, S.; Siddiqi, M. I. *Med. Chem. Res.* **2014**, *23* (7), 3308–3326.
- (66) Škedelj, V.; Perdih, A.; Brvar, M.; Kroflič, A.; Dubbée, V.; Savage, V.; O’Neill, A. J.; Solmajer, T.; Bešter-Rogač, M.; Blanot, D.; Hugonnet, J. E.; Magnet, S.; Arthur, M.; Mainardi, J. L.; Stojan, J.; Zega, A. *Eur. J. Med. Chem.* **2013**, *67*, 208–220.
- (67) Zhou, N.; Xu, Y.; Liu, X.; Wang, Y.; Peng, J.; Luo, X.; Zheng, M.; Chen, K.; Jiang, H. *Int. J. Mol. Sci.* **2015**, *16* (6), 13407–13426.
- (68) Sun, R.; Li, X.; Li, Y.; Zhang, X.; Li, X.; Li, X.; Shi, Z.; Bao, J. *J. Mol. Model.* **2015**, *21* (5), 133.
- (69) Pandey, R. K.; Kumbhar, B. V.; Srivastava, S.; Malik, R.; Sundar, S.; Kunwar, A.; Prajapati, V. K. *J. Biomol. Struct. Dyn.* **2017**, *35* (1), 141–158.

- (70) Larif, S.; Ben Salem, C.; Hmouda, H.; Bouraoui, K. *J. Mol. Graph. Model.* **2014**, *53*, 1–12.
- (71) Li, J.; Zhou, N.; Liu, W.; Li, J.; Feng, Y.; Wang, X.; Wu, C.; Bao, J. *J. Biomol. Struct. Dyn.* **2016**, *34* (5), 1101–1112.
- (72) Gangwal, R. P.; Damre, M. V.; Das, N. R.; Dhoke, G. V.; Bhadauriya, A.; Varikoti, R. A.; Sharma, S. S.; Sangamwar, A. T. *J. Mol. Graph. Model.* **2015**, *57*, 80–98.
- (73) Perdih, A.; Hrast, M.; Barreteau, H.; Gobec, S.; Wolber, G.; Solmajer, T. *Bioorganic Med. Chem.* **2014**, *22* (15), 4124–4134.
- (74) Bhadauriya, A.; Dhoke, G. V.; Gangwal, R. P.; Damre, M. V.; Sangamwar, A. T. *Mol. Divers.* **2013**, *17* (1), 139–149.
- (75) Barreca, M. L.; Iraci, N.; Manfroni, G.; Gaetani, R.; Guercini, C.; Sabatini, S.; Tabarrini, O.; Cecchetti, V. *J. Chem. Inf. Model.* **2014**, *54* (2), 481–497.
- (76) Xu, H.; Wang, Z.; Liang, X.; Li, X.; Shi, Z.; Zhou, N.; Bao, J. *Mol. Biosyst.* **2014**, *10* (6), 1524–1537.
- (77) Pathania, S.; Randhawa, V.; Bagler, G. *PLoS One* **2013**, *8* (4), e61327.
- (78) Gangwal, R. P.; Dhoke, G. V.; Damre, M. V.; Khandelwal, K.; Sangamwar, A. T. *J. Comput. Med.* **2013**, *2013*, 1–9.
- (79) Buchholz, T. Pflanzliche Sekundärprodukte als Inhibitoren von ausgewählten Verdauungsenzymen, 2013.
- (80) Ye, Y.; Zhang, B.; Mao, R.; Zhang, C.; Wang, Y.; Xing, J.; Liu, Y.-C.; Luo, X.; Ding, H.; Yang, Y.; Zhou, B.; Jiang, H.; Chen, K.; Luo, C.; Zheng, M. *Org. Biomol. Chem.* **2017**, *15* (17), 3648–3661.
- (81) Kesharwani, M.; Raghavan, S.; Gunasekaran, K.; Velmurugan, D. *J. Biomol. Struct. Dyn.* **2017**, 1–23.
- (82) Pandey, R. K.; Narula, A.; Naskar, M.; Srivastava, S.; Verma, P.; Malik, R.; Shah, P.; Prajapati, V. K. *J. Biomol. Struct. Dyn.* **2016**, 1–14.

- (83) Islam, M. A.; Pillay, T. S. *Mol. BioSyst.* **2016**, *12* (3), 982–993.
- (84) Manhas, A.; Lone, M. Y.; Jha, P. C. *J. Mol. Graph. Model.* **2017**.
- (85) Gurung, A. B.; Aguan, K.; Mitra, S.; Bhattacharjee, A. *J. Biomol. Struct. Dyn.* **2016**, *1102* (just-accepted), 1–40.
- (86) Mohammad Shahid. *Integrative Systems Approaches Towards Brain Pharmacology and Polypharmacology*, 2013.
- (87) LUO, Qi-Yao; WANG, Zi-Yun; JIN, Hong-Wei; LIU, Zhen-Ming; ZHANG, L.-R. *Acta Physico-Chimica Sin.* **2016**, *32* (10), 2606–2619(14).
- (88) Li, J.; Wang, H.; Li, J.; Bao, J.; Wu, C. *Int. J. Mol. Sci.* **2016**, *17* (7), 1055.
- (89) Garcia-Vallve, S.; Guasch, L.; Mulero, M. In *Foodinformatics*; Springer International Publishing: Cham, 2014; pp 151–176.
- (90) Colomina Perat, J. C. *Polifarmacología, desarrollo de una herramienta de “screening” de posibles nuevos fármacos activos para una determinada proteína “target”*, Universitat Oberta de Catalunya, 2016.
- (91) Murgueitio, M. S.; Wolber, G.; Daniela Schuster, P. *Targeting the first barrier in immune response: design of novel TLR2 antagonists and binding mode investigation*, 2013.
- (92) Pei, F.; Jin, H.; Zhou, X.; Xia, J.; Sun, L.; Liu, Z.; Zhang, L. *Chem. Biol. Drug Des.* **2015**, *86* (5), 1226–1241.
- (93) Wei, Y.; Li, J.; Chen, Z.; Wang, F.; Huang, W.; Hong, Z.; Lin, J. *Eur. J. Med. Chem.* **2015**, *101*, 409–418.
- (94) Husby, J.; Bottegoni, G.; Kufareva, I.; Abagyan, R.; Cavalli, A. *J. Chem. Inf. Model.* **2015**, *55* (5), 1062–1076.
- (95) Xia, J.; Jin, H.; Liu, Z.; Zhang, L.; Wang, X. S. *J. Chem. Inf. Model.* **2014**, *54* (5), 1433–1450.

- (96) Drwal, M. N.; Banerjee, P.; Dunkel, M.; Wettig, M. R.; Preissner, R. *Nucleic Acids Res.* **2014**, *42* (W1), W53–W58.
- (97) FU Haiyang , YAN Zhihui , LIN Runfang² , DIAO Aipo¹ , XIAO Dongguang. *J. Tianjin Univ. Sci. Technol.* **2014**, *29* (863), 2013–11.
- (98) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. *J. Med. Chem.* **2012**, *55* (14), 6582–6594.
- (99) Wang, L.; Pang, X.; Li, Y.; Zhang, Z.; Tan, W. *Bioinformatics* **2016**, *55* (8), btw783.
- (100) Braga, R.; Alves, V.; Silva, A.; Nascimento, M.; Silva, F.; Liao, L.; Andrade, C. *Curr. Top. Med. Chem.* **2014**, *14* (16), 1899–1912.
- (101) Yuriev, E.; Holien, J.; Ramsland, P. A. *J. Mol. Recognit.* **2015**, *28* (10), 581–604.
- (102) Chen, Y.-C. *Trends Pharmacol. Sci.* **2015**, *36* (2), 78–95.
- (103) Xia, J.; Tilahun, E. L.; Reid, T.-E. E.; Zhang, L.; Wang, X. S. *Methods* **2015**, *71* (C), 146–157.
- (104) Xia, J.; Tilahun, E. L.; Kebede, E. H.; Reid, T. E.; Zhang, L.; Wang, X. S. *J. Chem. Inf. Model.* **2015**, *55* (2), 374–388.
- (105) Braga, R. C.; Andrade, C. H. *Curr. Top. Med. Chem.* **2013**, *13* (9), 1127–1138.
- (106) Cruz-Monteagudo, M.; Schürer, S.; Tejera, E.; Pérez-Castillo, Y.; Medina-Franco, J. L.; Sánchez-Rodríguez, A.; Borges, F. *Drug Discov. Today* **2017**.
- (107) Pirhadi, S.; Sunseri, J.; Koes, D. R. *J. Mol. Graph. Model.* **2016**, *69*, 127–143.
- (108) Ambure, P.; Aher, R. B.; Roy, K. In *Computer-Aided Drug Discovery*; 2014; pp 257–296.

- (109) Garcia-Vallvé, S.; Guasch, L.; Tomas-Hernández, S.; Del Bas, J. M.; Ollendorff, V.; Arola, L.; Pujadas, G.; Mulero, M. *J. Med. Chem.* **2015**, *58* (14), 5381–5394.
- (110) Labbé, C. M.; Rey, J.; Lagorce, D.; Vavruša, M.; Becot, J.; Sperandio, O.; Villoutreix, B. O.; Tufféry, P.; Miteva, M. A. *Nucleic Acids Res.* **2015**, *43* (W1), W448–W454.
- (111) Deller, M. C.; Rupp, B. *J. Comput. Aided. Mol. Des.* **2015**, *29* (9), 817–836.
- (112) Peach, M. L.; Cachau, R. E.; Nicklaus, M. C. *J. Mol. Recognit.* **2017**, e2618.
- (113) van Beusekom, B.; Perrakis, A.; Joosten, R. P. *Methods Mol. Biol.* **2016**, *1415*, 107–138.
- (114) Emsley, P. *Acta Crystallogr. Sect. D Struct. Biol.* **2017**, *73* (3), 203–210.
- (115) Weichenberger, C. X.; Afonine, P. V.; Kantardjieff, K.; Rupp, B. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2015**, *71* (5), 1023–1038.
- (116) Lamb, A. L.; Kappock, T. J.; Silvaggi, N. R. *Biochim. Biophys. Acta - Proteins Proteomics* **2015**, *1854* (4), 258–268.
- (117) Touw, W. G.; Joosten, R. P.; Vriend, G. *J. Mol. Biol.* **2016**, *428* (6), 1375–1393.
- (118) Zheng, H.; Hou, J.; Zimmerman, M. D.; Wlodawer, A.; Minor, W. *Expert Opin. Drug Discov.* **2014**, *9* (2), 125–137.
- (119) Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M. *FEBS J.* **2013**, *280* (22), 5705–5736.
- (120) Grinter, S. Z.; Zou, X. *Molecules* **2014**, *19* (7), 10150–10176.
- (121) Touw, W. G.; Baakman, C.; Black, J.; Te Beek, T. A. H.; Krieger, E.; Joosten, R. P.; Vriend, G. *Nucleic Acids Res.* **2015**, *43* (D1), D364–D368.
- (122) Joosten, R. P.; Long, F.; Murshudov, G. N.; Perrakis, A. *IUCrJ* **2014**, *1* (4), 213–220.



UNIVERSITAT
ROVIRA i VIRGILI