Archives of Biochemistry and Biophysics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Archives of Biochemistry and Biophysics

journal homepage: www.elsevier.com/locate/yabbi

Review

Receptor-based virtual screening protocol for drug discovery

Nuno M.F.S.A. Cerqueira, Diana Gesto, Eduardo F. Oliveira, Diogo Santos-Martins, Natércia F. Brás, Sérgio F. Sousa, Pedro A. Fernandes, Maria J. Ramos *

UCIBIO, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

ARTICLE INFO

Article history:
Received 2 March 2015
and in revised form 26 May 2015
Available online xxxx

Keywords:
Virtual screening
Molecular docking
Scoring functions
Search algorithms
Drug discovery

ABSTRACT

Computational aided drug design (CADD) is presently a key component in the process of drug discovery and development as it offers great promise to drastically reduce cost and time requirements.

In the pharmaceutical arena, virtual screening is normally regarded as the top CADD tool to screen large libraries of chemical structures and reduce them to a key set of likely drug candidates regarding a specific protein target. This chapter provides a comprehensive overview of the receptor-based virtual screening process and of its importance in the present drug discovery and development paradigm. Following a focused contextualization on the subject, the main stages of a virtual screening campaign, including its strengths and limitations, are the subject of particular attention in this review. In all of these stages special consideration will be given to practical issues that are normally the Achilles heel of the virtual screening process.

Introduction

The process of drug discovery is very complex and requires an interdisciplinary effort to design effective and commercially feasible drugs. The objective of drug design is to find a drug that can interact with a specific drug target and modify its activity. The drug targets are generally proteins that perform most of the tasks needed to keep cells alive. Drugs are small molecules that bind to a specific region of a protein and can turn it on or off. Some very powerful drugs, such as antibiotics or anticancer drugs, are used to completely disable a critical protein in the cell. These drugs can kill bacteria or cancer cells.

It is generally recognized that drug discovery and development are very time and resource-consuming processes and the whole process is often compared to searching for a needle in a haystack. It is estimated that a typical drug discovery cycle, from lead identification to clinical trials, can take 17 years with a cost of 800 million US dollars. In this process it is estimated that five out of 40,000 compounds tested in animals eventually reach human testing and only one in five compounds that enter clinical studies is approved. This represents an enormous investment in terms of time, money and human resources. It includes chemical synthesis, purchase, and biological screening of hundreds of thousands of compounds to identify hits followed by their optimization to generate leads, which require further synthesis. In addition, predictability of animal studies in terms of both efficacy and toxicity is frequently suboptimal. Therefore, new approaches are needed to facilitate, expedite and streamline drug discovery and development, save time, money and resources.

On October 5, 1981, Fortune magazine published a cover article entitled "Next Industrial Revolution: Designing Drugs by Computer at Merck". Some have credited this as being the start of intense interest in computer-aided drug design (CADD)[1] [1].

CADD is defined by the IUPAC as all computer assisted techniques used to discover, design and optimize compounds with desired structure and properties. CADD has emerged from recent advances in computational chemistry and computer technology, and promises to revolutionize the design of functional molecules. The ultimate goal of CADD is to virtually screen a large database of compounds to generate a set of hit compounds (active drug candidates), lead compounds (most likely candidates for further evaluation), or optimize known lead compounds, *i.e.* transform biologically active compounds into suitable drugs by improving their physicochemical, pharmaceutical and ADMET/PK (pharmacokinetic) properties [2].

* Corresponding author.
E-mail address: mjramos@fc.up.pt (M.J. Ramos).

[1] *Abbreviations used:* CADD, computer-aided drug design; VS, virtual screening; MC, Monte Carlo; GA, Genetic Algorithms; RMSD, Root Mean Square Deviation; MD, Molecular Dynamics; FEP, Free Energy Perturbation; TI, Thermodynamic Integration; PLP, Piecewise Linear Potential; PMF, Potential of Mean Force; EF, enrichment factor; ROC, Receiver Operator Characteristic; TPR, true positive rate; FPR, false positive rate.

http://dx.doi.org/10.1016/j.abb.2015.05.011
0003-9861/

The fast expansion and popularity of this field of research has been made possible partially by the advances in software and hardware, computational power and sophistication. On the other hand, the knowledge of the 3D shapes of proteins, nucleic acids, and complex assemblies are fundamental to understand all aspects of potential drug targets. It is remarkable that, from 1970 to 2004, 50,000 structures have been deposited on the protein databank, in 2014 this number has tripled to 150,000 and in 2018 it is expected that this latter number doubles. In addition, the increasing digital repositories containing detailed information on potential drugs and other useful compounds provide goldmines for the design of new drugs.

CADD is widely used in the pharmaceutical industry to improve the efficiency of the drug discovery and development pipeline. One method that was quickly adopted was the virtual screening of large compound databases against drug targets. The goal is to select a set of molecules with desirable properties (active, drug-like, lead-like) targeting a specific protein and eliminate compounds with undesirable properties (inactive, reactive, toxic, poor ADMET/PK). The computational methodologies used for this purpose are known as virtual screening methodologies.

The generic definition of virtual screening encompasses many different methodologies, which are generally divided in two main classes: the ligand-based virtual screening methods and the receptor-based virtual screening methods.

Ligand-based virtual screening methods aim to identify molecules sharing common features, both at the chemical and physical levels grounded in the assumption that similar compounds can have similar effects on a drug target [3]. These methods normally discard all information related to the drug target and focus exclusively on the ligand. Within the lock-and-key paradigm, these approaches compare different keys, and neglect the lock. Thus, the model of the receptor is only implicitly built based on what binds to it [4]. The main downside of these methods is that substantial activity data regarding the compounds that are studied are required to get reasonable results.

Receptor-based virtual screening methods, also called structure-based methods, require the existence of a 3D structure of the target. These methods involve explicit molecular docking of each ligand into the binding site of the target, producing a predicted binding mode for each database compound, together with a measure of the quality of the fit of the compound in the target-binding site. This information is then used to sort out ligands that bind strongly to the target protein from ligands that do not. Receptor-based approaches are gaining considerable importance over ligand-based techniques, particularly as more and more 3D structures of target proteins are determined and become available, and also because the results tend to be more reliable and accurate. The current state-of-the-art of receptor-based virtual screening is reviewed in this chapter, and general approaches, successes and pitfalls associated with the technology are highlighted.

The screening process

Receptor-based virtual screening encompasses a variety of sequential computational stages, including target and database preparation, docking and post-docking analysis, and prioritization of compounds for experimental testing. A typical workflow of a receptor-based virtual screening is presented in Fig. 1. All stages of this workflow depend on sound implementation of a wide range of computational techniques that will be discussed in detail in the following sections. In each section special attention will be given to practical issues that are normally the Achilles heel of the virtual screening process. Since in this book chapter only the receptor-based virtual screening will be reviewed, we are going to adopt the general term, virtual screening (VS), to describe this type of screening methodologies.

Target selection

Target selection is among the first stages of a virtual screening campaign and it is pivotal for a successful drug development process. Among the four types of macromolecules that can be targeted (proteins, polysaccharides, lipids and nucleic acids) with small-molecule compounds, proteins, and within those enzymes, are generally the first choice, since their binding pocket properties allow for high specificity, potency and low toxicity. When considering a potential protein target to modify a disease it should be pondered if it is advantageous to select an upstream, widely implicated target or, instead, a downstream target, very specific to the pathway that we want to tackle.

Once a protein target with the potential to modify the disease has been identified, it is time to get its 3D structure. The Protein Data Bank is the leading repository for experimentally determined 3D structures of large biological molecules. This database is therefore the first approach to retrieve a protein 3D structure for a VS campaign. In the case that the experimental 3D structure of the protein does not exist, then homology-modeling methods can be used to build it. There are several examples in the literature showing that these homology models can be used with success in VS campaigns [5–7].

Binding site detection

Once the 3D structure of a protein has been obtained then it is possible to evaluate its *druggability* score. The *Druggability* can be understood as the capability that a receptor has to bind molecules with drug like properties. This depends of course on the ability of the molecule to favorably interact with a particular pocket or cleft in that protein. The location of these binding sites is easy when a ligand has been co-crystallized explicitly with the target protein. However, when this sort of information is not available, the location of the binding site can be cumbersome. In these cases computational tools can be used to identify and characterize potential binding sites. Among the available computational tools, some algorithms rely mostly on geometric characteristics to search for binding pockets, such as POCKET [8], LIGSITE [9], SURFNET [10], SPHGEN [11], FPOCKET [12], etc., while others, such as Q-SITEFINDER [13], GRID [14,15], POCKETPICKER [16], FLAPSITE [17], CS-MAP algorithm [18], in order to calculate the energy of probes interacting with potential binding sites to identify and rank them. ***Geometry-based algorithms*** are usually prized because they are fast and robust in dealing with structural variations or missing atoms/residues in the input structure [19]. ***Energy-based algorithms***, on the other hand, are often more sensitive and specific [20]. Despite the distinctive approaches, the performance is very similar and both methods can correctly predict 95% of the known binding sites [21].

Target preparation

After the target has been defined and the most *druggable* binding site chosen, it is necessary to prepare the target for docking. The general steps in target preparation require removing solvent and ligand molecules, adding hydrogen atoms, setting up bond orders and formal charges, capping chain termini and defining amino acid protonation states (atom types). It might also be necessary to refine the crystallographic structure and define the binding site portions that will be left flexible. Target preparation is usually

Please cite this article in press as: N.M.F.S.A. Cerqueira et al., Arch. Biochem. Biophys. (2015), http://dx.doi.org/10.1016/j.abb.2015.05.011