Scoring Techniques in Virtual Screening A Review

Diego Garcia

May 3, 2018

1 Introduction

Historically, drug discovery has been expensive, time consuming, and inaccessible to academia [1, 7]. These challenges are due to the process that is endured from the creation of the drug up until when and if the drug passes all clinical testing and is approved by the FDA [1]. This process starts with the screening of compounds to gather information about them. Until recently, drugs were discovered by screening all possible chemical-poses limited by the compounds available in a lab or chemical collection [5]. Methods have since been developed that do not require obtaining each compound but rather use a virtual model of each compound. One such method is virtual screening [1–7].

Virtual Screening has two different variations: ligand-based virtual screening and structure-based virtual screening. Ligand-based virtual screening involves detailed analysis of ligands to predict how similar ligands will react. Structure-based involves the analysis of a compound in which the 3D structure has been identified [7].

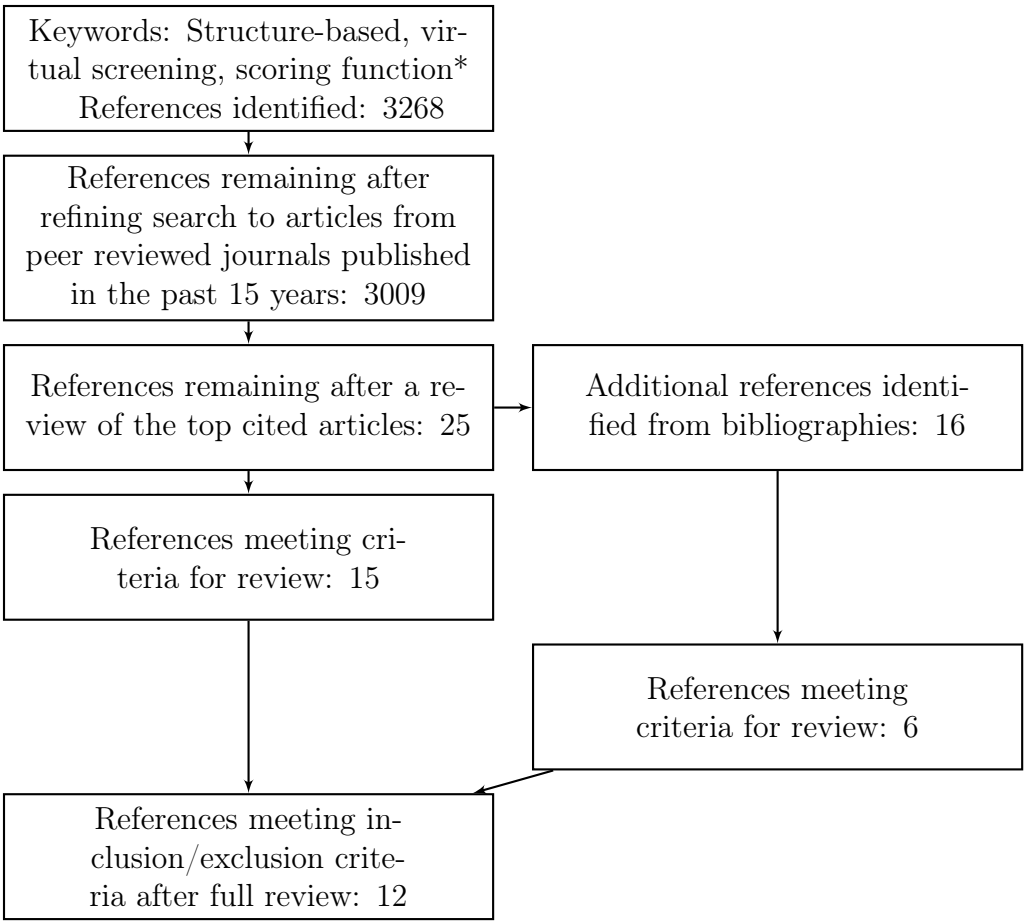
In structure-based virtual screening computational methods are used to test thousands of possible interactions between a target receptor and compounds in a chemical library. The quality of the interactions are formulated into a mathematical function called a scoring function [7]. Creating scoring functions is a crucial aspect of structure-based virtual screening [3].

The purpose of this paper is to analyze the different scoring methods and to describe which methods are considered best in practice and why.

2 Methods

This research for this review was conducted using the University of Miami Library System. A search, limited to articles from peer reviewed journals, was made. The search results were filtered through an inclusion exclusion criteria detailed in **Figure 1**. The remaining articles were examined in detail and grouped according to subject similarities.

Figure 1: Search conducted via University of Miami Library System



3 Results

The goal of scoring functions is to predict binding energies or to rank compounds in order of their reactivity [6]. They are a crucial aspect in virtual screening because they ultimately determine the accuracy of the algorithm. Thus, the design of reliable scoring functions is vital.

There are several different methods to create scoring functions but it is commonly stated that each can be separated into three categories: force-field based, empirical, and knowledge-based [1–3, 5, 6]. Some also propose the idea of other categories of scoring functions, such as one called descriptor-based scoring [4]. Each of these techniques analyze protein-ligand complexes and produce a scoring function based on different features.

3.1 Forcefield-based

Forcefield-based scoring functions are based on the summation of protein-ligand interaction energies and internal ligand energies using well-documented molecular mechanics forcefield parameters [1–4, 6]. These terms include van der Waals forces and electrostatic energies.

Many force-field based scoring functions only consider a single protein conformation which greatly simplifies scoring. These scoring functions will also improve as advances are made in quantum mechanics and other high-level physics theories [4].

A major weakness of forcefield-based scoring functions is that they do not account for solvation and entropic terms natively [2,3]. The use of forcefields also sometimes requires a cut-off distance which complicates the accurate reading of some binding effects [1, 3]. They are also known to score larger ligands better because of larger size [1]

3.2 Empirically-derived

Empirically-derived scoring functions are derived by summing a set of individual energy terms, terms such as hydrogen bonds, hydrophobic effects, entropy, etc. [2, 4, 6]. The coefficients of theses terms are obtained using regression analysis of experimentally determined binding energies and structural information [3].

A strength of empirically-derived scoring functions is that the terms are often simple to evaluate. Another strength is that customized scoring functions can be developed by adding or omitting any specific energy term [4]. Doing so can help develop scoring functions specific to a chosen set of energy terms which might be helpful for specific target ligands.

A weakness of empirically-derived scoring functions is their dependency on molecular data sets (training sets) to perform their regression analyses [2,3]. This means that the applicability of these scoring functions are dependent on these training sets. Although this is the case, empirically-derived scoring functions will continue to improve with the rapid increase in the number of known protein-ligand structures and the improvement of commonly used training sets.

3.3 Knowledge-based

In knowledge-based scoring functions energy terms are derived directly from experimentally determined structural information, usually from a protein database such as PDB. Then, binding energy is calculated by summing the interaction terms of each possible protein-ligand atom pair in the entire complex using Boltzmann distributions [1,2].

Knowledge-based scoring functions are computationally efficient and simplistic [1, 3] This permits the efficient screening of large compound databases relatively quickly [3].

A major weakness is apparent in knowledge-based scoring functions, underrepresented atom interactions in protein databases are not scored [1]. These scoring functions do not account for certain terms such as hydrogen bonding terms and solvation effects, therefore additional terms often need to be added [2].

3.4 Other

Recently there has been evidence of a type of scoring function that does not fall into one of the aforementioned categories. These methods introduce new developments in the field of computational intelligence [1, 4]. They employ machine learning algorithms to learn using a training set of protein-ligand descriptors, such as electrostatic energies and hydrogen bonds, and derive their final model from it.

It has been shown that these methods have achieved better scores than any other type of scoring function [4]. A recent study showed that a specific computational intelligence strategy—deep learning—improved scoring functions and produced some of the best reported results so far on a commonly used data set [7].

4 Discussion

Moitessier et al. compared the accuracy of a number of Forcefield-based, empirically-derived and knowledge-based scoring functions currently in use in docking programs today. They found that none of the assessed programs were consistently better than others and that most exhibited similar accuracies [6]. This shows that none of the three main scoring categories is better than the other. Commonly used docking programs make assumptions and simplifications in the evaluation of scoring functions and do not fully account for a number of terms [3]. Many times in practice multiple scoring functions are combined in an effort to better account for certain terms and to try to balance errors made by individual scoring functions [2]. The combination of information from multiple scoring functions is considered to be called consensus scoring [1–3].

Further advances in respect to scoring functions seem to be coming from machine learning techniques and from consensus scoring. As new computational intelligence techniques arise, better machine learning techniques will be created and as new advances in any of the three major scoring categories arise, Consensus scoring will improve.

References

[1] D. Hecht and G. Fogel, *Computational intelligence methods for docking scores*, Current Computer-Aided Drug Design, 5 (2009), pp. 56–68.

[2] S. Huang and X. Zou, *Advances and challenges in protein-ligand docking*, Int. J. Mol. Sci., (2010), pp. 3016–3034.

[3] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, *Docking and scoring in virtual screening for drug discovery: Methods and applications*, Nature Reviews Drug Discovery, (2004), pp. 935–949.

[4] J. Liu and R. Wang, *Classification of current scoring functions*, Journal of Chemical Information and Modeling, (2015), pp. 475–482.

[5] C. McInnes, *Virtual screening strategies in drug discovery*, Current Opinion in Chemical Biology, (2007), pp. 494–502.

[6] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. Corbeil, *Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go*, British Journal of Pharmacology, (2008).

[7] J. C. Pereira, E. R. Caffarena, and C. N. dos Santos, *Boosting docking-based virtual screening with deep learning*, Journal of Chemical Information and Modeling, (2016), pp. 2495–2506.

[8] M. Wójcikowski, P. J. Ballester, and P. Siedlecki, *Performance of machine-learning scoring functions in structure-based virtual screening*, Scientific Reports, 7 (2017).