Methods 71 (2015) 77–84

Contents lists available at ScienceDirect

Methods

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

Protein structure prediction provides comparable performance to crystallographic structures in docking-based virtual screening

Hongying Du [a,b], Jeffrey R. Brender [a], Jian Zhang [a], Yang Zhang [a,*]

[a] Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA
[b] Department of Public Health, Lanzhou University, Lanzhou 730000, China

ARTICLE INFO

Article history:
Available online 8 September 2014

Keywords:
Virtual screening
Enrichment rate
Ligand docking
Protein structure prediction

ABSTRACT

Structure based virtual screening has largely been limited to protein targets for which either an experimental structure is available or a strongly homologous template exists so that a high-resolution model can be constructed. The performance of state of the art protein structure predictions in virtual screening in systems where only weakly homologous templates are available is largely untested. Using the challenging DUD database of structural decoys, we show here that even using templates with only weak sequence homology (<30% sequence identity) structural models can be constructed by I-TASSER which achieve comparable enrichment rates to using the experimental bound crystal structure in the majority of the cases studied. For 65% of the targets, the I-TASSER models, which are constructed essentially in the apo conformations, reached 70% of the virtual screening performance of using the holo-crystal structures. A correlation was observed between the success of I-TASSER in modeling the global fold and local structures in the binding pockets of the proteins versus the relative success in virtual screening. The virtual screening performance can be further improved by the recognition of chemical features of the ligand compounds. These results suggest that the combination of structure-based docking and advanced protein structure modeling methods should be a valuable approach to the large-scale drug screening and discovery studies, especially for the proteins lacking crystallographic structures.

1. Introduction

Virtual screening is a computational approach to detect potential leads from compound libraries that has become a standard technology in modern drug discovery pipelines [1]. The total number of potential ligands for drug development is much larger than what can be feasibly tested. While estimates of the total number of synthetically accessible small molecules vary, even the smallest number indicates a drug-like chemical space that is much larger than what can be efficiently explored experimentally through blind screening. Given the common estimate that a single industrial lab can only test 10,000–100,000 compounds in a day with standard high throughput screening, the smallest estimate [2] of drug-like chemical molecules (1.5×10^7) still presents a formidable task for lead selection. If larger estimates of 10^{23}–10^{60} possible drug-like molecules are considered [3], the total number of potential ligands for drug development is much larger than what can be feasibly tested experimentally. The main goal of virtual screening is therefore to identify a limited set of candidates to be synthesized for the much more expensive next step of experimentally examining their biological activities [1].

Historically, virtual screening approaches in the drug development process have been divided into structure- and ligand-based algorithms [4,5]. Structure-based computational modeling approaches such as molecular docking use the full three dimensional structure of the protein target for lead optimization and hit discovery [6]. The ligand-based approach, by contrast, ignores the structural details of the protein target and finds ligands with pharmacophores similar to known hits to generate a model of the pharmacodynamics of a potential hit, or to perform quantitative structure–activity relationship studies [5]. In principle, the structure-based methods might be expected to give better results than the ligand-based approaches, because they try to simulate the intrinsic character of protein–ligand interactions [7]. However, a major drawback of the structure-based technique is a structural model of the protein, which usually needs to have high-resolution, must be available, which is frequently not the case for many protein families of interest in drug development. If a high-resolution structural model cannot be created, only ligand-based approaches may be used.

* Corresponding author. Fax: +1 734 615 6553.
E-mail address: zhng@umich.edu (Y. Zhang).

http://dx.doi.org/10.1016/j.ymeth.2014.08.017
1046-2023/

Although the amount of high-resolution protein structures has increased dramatically in recent years, the structures of some important protein targets implicated in the etiology of deadly diseases remain unsolved [8,9]. What can be done if the 3D protein structure of the drug target is not available? Fortunately, many computational methods have successfully predicted accurate 3D structures from only the amino-acid sequence of the target. Several methods have been used for protein structure prediction including homology modeling [10,11], threading [12,13], and *ab initio* folding [14–16].

Most virtual screening studies using predicted structures have been relied on homology modeling, which is based on the general observation that proteins with similar sequences can be expected to possess similar structures. Homology modeling of proteins consists of identification of related proteins with a known 3D structure that can serve as a template, followed by sequence alignment of the target and template, and the refinement of the structural model. Although there are specific cases where a template with low sequence similarity may adopt similar structure folds (e.g. 27 different homologous subfamilies from 60 different enzyme classifications, which have no sequence similarity, have the same TIM barrel fold [17]), homologous templates generally refers to a known protein that shares strong sequence similarity to the target. Thus, the final quality of a homology model for virtual screening often depends on the level of sequence identity between the target and template. Multiple studies have attempted to assess the degree of sequence identity needed for effective virtual screening for different classes of protein targets. As an approximate rule, $\geqslant$50% sequence identity is believed to be sufficient for drug discovery [18–20], although this number varies widely among the target class and a strong correlation between sequence identity of the template and virtual screening success has not been verified for most targets at high sequence identity levels [21,22]. On the other hand, the accuracy of the structural model has been shown to correlate with virtual screening success [23]. The accuracy of homology modeling significantly declines when a template above 30% sequence identity cannot be found.

However, approaches based on advanced algorithms including threading and *ab initio* folding can increase the success rate for modeling the structure of distantly- or non-homologous protein targets [24]. The Iterative threading assembly refinement (I-TASSER) is one of such approaches that was designs to combine multiple pipelines of threading, *ab initio* folding and atomic-level structure refinement for full-length protein structure prediction [25]. In the recent community-wide blind structure prediction experiments, the Critical Assessment of Structure Prediction (CASP), I-TASSER has shown advantages over peer modeling programs in automated 3D structure predictions [26–30].

In this work, we tested the use of the I-TASSER models in large-scale structure-based virtual screening of the Directory of Useful Decoys (DUD) database [31]. The 3D structures of protein targets from the DUD database are first constructed by the I-TASSER program from the amino acid sequence alone, where template structures with a sequence identity >30% were excluded from the threading library. Next, atomic level refinement is performed by fragment guided molecular dynamics, FG-MD [32], to relax the predicted structures. The actual virtual screening is performed by molecular docking using the GRID score of DOCK 6.3 [33,34] to measure the binding site complementarity. While the performance of virtual screening using I-TASSER models did not match that of virtual screening using the experimental crystal structure, good enrichment rates ($\sim$70%) relative to using crystal structures could be achieved in most cases (65% of the structures tested) using the automatic structure prediction and docking pipelines without human intervention. The rate of success correlates well with the accuracy of I-TASSER in predicting the global fold and local structure of the binding pockets of the proteins. These results suggest that 3D models built by the state of the art structure prediction methods can provide a useful starting point of structure based virtual screening for the many cases where neither an experimental structure nor a clearly homologous template is available.

2. Materials and methods

2.1. Target set of proteins and ligands for virtual screening

We used the Directory of Useful Decoys (DUD) [31], one of the largest freely available databases for evaluating docking based virtual screening methods, to benchmark the performance of both crystal structure and I-TASSER predicted model based virtual screening. The DUD database consists of 40 protein targets from the Protein Data Bank (PDB). For each protein target, there are on average 74 active compounds (or 2950 active compounds in total), where for each active compound there are on average 36 inactive compounds (called decoys) with similar physical properties to the active compound but with dissimilar chemical topology [31]. Three out of the forty proteins in the DUD target set, including HIV-PR (1hpx), FXa (1f0r), HMGR (1hw8), are multi-chain proteins, the models of which should be constructed by the combination of I-TASSER with quaternary structure modeling tools [35]. Since the focus of this study is on automatic I-TASSER-based modeling and docking, these three proteins were removed from the test set. Finally, a crystal structure is not available for the kinase PDGFrb making a comparison impossible. The 36 remaining proteins are listed in Table 1, along with the PDB codes of the proteins and the number of actives and decoys for each target. In this study, only the decoys associated with a target were docked to that target (DUD-self), rather than all decoys for all targets.

Crystallographic structures of the bound proteins were used without further refinement after removing water and heavy metal atoms and adding polar hydrogens with ANTECHAMBER [36]. AM1-BCC partial charges [37,38] were added to both the crystallographic structures and I-TASSER models with ANTECHAMBER.

2.2. Creation of protein models by I-TASSER

The predicted structure models used for virtual screening were generated by the automated I-TASSER pipeline [27]. While the I-TASSER method has been described in previous work [17,20], we give an outline of the pipeline below.

In the first step of the I-TASSER modeling, the target sequences are threaded by LOMETS [39], a locally installed meta-server platform consisting of 8 threading proteins (FFAS [40], HHsearch [41], MUSTER [42], PPA [43], PRC [44], PROSPECT2 [45], SAM-T02 [46], SP3 [47], and SPARKS [48]), through a representative PDB library to search for possible folds or super-secondary structure segments matching the target sequence. In this benchmark test, all templates with a sequence identity >30% to the target are excluded to filter out homology contaminants. This cutoff corresponds to the "twilight zone" where structure prediction becomes significantly more difficult and therefore represents a challenging test where conventional homology modeling frequently fails [49].

Following the template detections, continuous fragments are excised from the LOMETS alignments, which are used to reassemble the full-length structure models. The threading unaligned regions (mainly loops and tails) are built by *ab initio* folding based on an on-lattice system. The structural assembly procedure is implemented by the replica-exchange Monte Carlo simulation [50], with an optimized knowledge-based force field. The models with the lowest free-energy are identified by SPICKER that clusters all structure decoys in the MC simulations [51].