Journal of Structural Biology 173 (2011) 558–569

Contents lists available at ScienceDirect

Journal of Structural Biology

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

The utility of geometrical and chemical restraint information extracted from predicted ligand-binding sites in protein structure refinement

Michal Brylinski, Seung Yup Lee, Hongyi Zhou, Jeffrey Skolnick *

Center for the Study of Systems Biology, Georgia Institute of Technology, Atlanta, GA 30318, United States

ARTICLE INFO

Article history:
Available online 17 September 2010

Keywords:
Ligand-binding site refinement
Protein threading
Protein structure prediction
Ligand-binding site prediction
Ensemble docking
Molecular function

ABSTRACT

Exhaustive exploration of molecular interactions at the level of complete proteomes requires efficient and reliable computational approaches to protein function inference. Ligand docking and ranking techniques show considerable promise in their ability to quantify the interactions between proteins and small molecules. Despite the advances in the development of docking approaches and scoring functions, the genome-wide application of many ligand docking/screening algorithms is limited by the quality of the binding sites in theoretical receptor models constructed by protein structure prediction. In this study, we describe a new template-based method for the local refinement of ligand-binding regions in protein models using remotely related templates identified by threading. We designed a Support Vector Regression (SVR) model that selects correct binding site geometries in a large ensemble of multiple receptor conformations. The SVR model employs several scoring functions that impose geometrical restraints on the Cα positions, account for the specific chemical environment within a binding site and optimize the interactions with putative ligands. The SVR score is well correlated with the RMSD from the native structure; in 47% (70%) of the cases, the Pearson's correlation coefficient is >0.5 (>0.3). When applied to weakly homologous models, the average heavy atom, local RMSD from the native structure of the top-ranked (best of top five) binding site geometries is 3.1 Å (2.9 Å) for roughly half of the targets; this represents a 0.1 (0.3) Å average improvement over the original predicted structure. Focusing on the subset of strongly conserved residues, the average heavy atom RMSD is 2.6 Å (2.3 Å). Furthermore, we estimate the upper bound of template-based binding site refinement using only weakly related proteins to be ~2.6 Å RMSD. This value also corresponds to the plasticity of the ligand-binding regions in distant homologues. The *B*inding *S*ite *R*efinement (BSR) approach is available to the scientific community as a web server that can be accessed at http://cssb.biology.gatech.edu/bsr/.

1. Introduction

With the rapid accumulation of protein sequences generated by the now numerous genome-sequencing projects (Aury et al., 2008; Tettelin and Feldblyum, 2009; Wheeler et al., 2008), the key challenge in biological sciences has shifted from the study of single molecules to the exhaustive exploration of molecular interactions and biological processes at the level of complete proteomes (Butcher et al., 2004; You, 2004). To achieve the ambitious goal of characterizing and understanding the molecular function of all gene products in a given proteome, a number of structure-based approaches to protein function inference have been developed (Juncker et al., 2009; Loewenstein et al., 2009; Rost et al., 2003). Contemporary methods for binding site detection are fairly insensitive to the overall quality of the target structures (Brylinski and Skolnick, 2008a) and facilitate the selection of correctly predicted models in protein structure prediction (Chelliah and Taylor, 2008). Approximate protein models can be routinely generated by the state-of-the-art structure prediction techniques for the majority of gene products in a given proteome (Fiser, 2004; Gopal et al., 2001; Yura et al., 2006; Zhang and Skolnick, 2004a); this opens up the possibility of using low-to-moderate resolution models for genome-wide function annotation.

Qualitative protein function annotation using Enzyme Commission (EC) numbers or Gene Ontology (Ashburner et al., 2000) terms is typically followed by a comprehensive functional characterization at the molecular level. The studies of interactions between proteins and other molecular species in a cell are routinely supported by computations involving docking of DNA (Gao and Skolnick, 2009; van Dijk and Bonvin, 2008), other protein partners (Lyskov and Gray, 2008; Wiehe et al., 2008) and small ligands (Goodsell et al., 1996; Moustakas et al., 2006). In the latter case, the docking of specific ligands can be extended to large-scale

* Corresponding author.
E-mail address: skolnick@gatech.edu (J. Skolnick).

1047-8477/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2010.09.009

virtual screening of combinatorial libraries in order to discover novel bioactive compounds (Rajamani and Good, 2007; Seifert et al., 2007). Notwithstanding the advances in the development of docking approaches and scoring functions, the application of many ligand docking/screening algorithms to protein models is limited by the quality of the binding site in the target structure; mean structure rearrangements greater than 1.5 Å may cause the loss of even 90% of the docking accuracy (Erickson et al., 2004). Many other benchmark studies report a notable drop-off in the docking accuracy when non-native structures are used as the target receptors (Murray et al., 1999; Sutherland et al., 2007; Wu et al., 2003).

Despite progress in protein structure prediction (Kryshtafovych et al., 2005), theoretical models, particularly those modeled using remote homology, still have significant structural inaccuracies in ligand-binding sites (DeWeese-Scott and Moult, 2004; Piedra et al., 2008); this has stimulated the development of methods for the local refinement of binding pocket residues prior to ligand docking. The local refinement of ligand-binding regions is complicated by many factors. The conformational changes triggered by ligand binding may require side chain geometries (Heringa and Argos, 1999) absent in standard rotamer libraries (Dunbrack and Karplus, 1993; Koehl and Delarue, 1994). Moreover, it has been demonstrated that there is no correlation between the backbone movement of a residue upon binding and the flexibility of its side chain (Najmanovich et al., 2000). To tackle the difficult problem of binding site modeling, Kauffman and colleagues incorporated information on the residues involved in ligand binding in constructing the target-template alignments and observed an improvement in the overall quality of the modeled ligand-binding regions (Kauffman et al., 2008). In principle, ligand molecules could also be explicitly used to model the binding sites. However, due to imperfections of available all-atom force fields, inclusion of protein flexibility in ligand docking against non-native receptor structures typically does not the improve root-mean-square deviation, RMSD of the binding pocket residues from the native structure (Davis and Baker, 2009). A slightly different approach, MOBILE, includes information about bioactive molecules as spatial knowledge-based restraints in the iterative refinement of protein models constructed using close homology (Evers et al., 2003). The issue is what happens when no closely related homologous structures are solved for the protein target of interest.

In this study, we describe a new template-based approach to the local refinement of ligand-binding regions in protein models that exploits the information provided by remotely related templates. We begin with an analysis of the plasticity of ligand-binding regions in distant homologues which provides an estimate of what would be the upper bound for the template-based refinement accuracy using only weakly related binding pockets. This also provides interesting insights into how structurally degenerate are similar/identical binding geometries in nature. Building on the resulting insights, we propose a new ligand-binding site refinement procedure that consists of the following: first, a large ensemble of multiple receptor conformations is generated. Then, a fitness function is applied to rank the structurally diverse set of constructed binding site geometries. This function comprises four scoring terms, whose parameters are derived from weakly related templates identified by threading (Jones and Hadley, 2000). The individual terms provide geometrical restraints on the Cα positions and Cα–Cα distances, account for a specific chemical environment within a binding site and optimize the interactions with putative ligands. The scoring functions are used to train a Support Vector Regression model to rank multiple receptor conformations. Here, for a large benchmark set, we apply this model to refine ligand-binding regions in proteins that are weakly homologous to their closest template whose structure is known and show that the SVR-based ranking selects fairly good binding site geometries. The *B*inding *S*ite *R*efinement (BSR) approach presented in this paper is available to the scientific community as a web server that can be accessed at http://cssb.biology.gatech.edu/BSR/.

2. Materials and methods

2.1. Dataset

Protein–ligand complexes used in this study were taken from the Protein-Small-Molecule Database (PSMDB) (Wallach and Lilien, 2009), a non-redundant repository of small molecule complexes for protein–ligand interaction studies. We selected proteins up to 200 residues in length, for which at least three weakly homologous (<35% sequence identity) template structures can be identified by threading (Skolnick and Kihara, 2001; Skolnick et al., 2004; Zhou and Zhou, 2004, 2005). Furthermore, we excluded those proteins that bind very small (<6 heavy atoms) as well as very big (>100 heavy atoms) ligands. The total number of complexes in the dataset is 904. Finally, we used only those targets for which the binding site center of mass can be predicted by FINDSITE within a distance of 6 Å. Since the accuracy of binding site prediction depends on the quality of the target structure, the number of proteins used for binding site refinement ranges from 662 for crystal structures to 440 for the most distorted models with an average RMSD (root-mean-square deviation) from the crystal structure of 9 Å; see additional details below. The PDB identifiers for the dataset proteins are provided in Supplementary materials, SI Table 1. Moreover, the entire dataset as well as the modeling results are available from http://cssb.biology.gatech.edu/BSR/.

2.2. All-atom RMSD of similar binding pockets

Due to significant sequence variability in remotely related proteins, the RMSD is typically calculated over Cα atoms. Here, we develop a simple method to calculate the heavy atom RMSD of similar, but not identical pockets extracted from weakly homologous template complexes. Residue equivalences are obtained from global structure alignments by fr-TMalign (Pandit and Skolnick, 2008; Zhang and Skolnick, 2005a), whereas the equivalent atoms in residue side chains are calculated by SMSD (Small Molecule Subgraph Detector) (Rahman et al., 2009). SMSD is a graph-based algorithm developed to identify the exact atom–bond equivalence between the query and target organic molecules in chemical similarity searches. Here, we apply SMSD to match the heavy atoms of different residue side chains. The all-atom RMSD calculated over the atoms matched for all binding residue pairs within a common pocket is denoted as $RMSD^{res}$. For a given pocket, ligand-binding residues can be divided into three groups, depending on the conservation of their binding patterns in evolutionarily related proteins. Strongly, moderately and weakly conserved binding residues are defined based on the fraction of templates that have a residue in an equivalent position in contact with a ligand: >0.75, 0.50–0.75, and 0.25–0.50, respectively. $RMSD^{res}$ values calculated over strongly, moderately and weakly conserved binding residues are denoted as $RMSD^{res}_{0.75}$, $RMSD^{res}_{0.50}$ and $RMSD^{res}_{0.25}$, respectively. In the RMSD calculations for the ligand-binding regions, we can also include the coordinates of bound ligands. Again, we use SMSD to establish the atom equivalences in ligand structures; the combined RMSD calculated over the heavy atoms of both protein residues and ligands is denoted as $RMSD^{res+lig}$.

2.3. Protein structure modeling

For each protein, we have constructed several models with different accuracy in terms of their RMSD and TM-score (Zhang and Skolnick, 2004b) from the native structure. In addition to the