FRAGMENT BINDING PREDICTION USING UNSUPERVISED LEARNING OF LIGAND SUBSTRUCTURE BINDING SITES

Grace Tang[1] *& Russ Altman*[1,2,*]

Depts. of Bioengineering[1] *and Genetics*[2]*, Stanford University.* [*]*russ.altman@stanford.edu*

Structure-based methods predicting small molecule binding focus on large drug-like molecules and have limited ability to predict fragment binding. Fragments, however, can be a key starting point in drug design and screening for a protein target. We leverage the redundancy among small molecules at the substructure level to develop a general structure-based predictor for fragment binding.

INTRODUCTION

Fragment-based drug design focuses on optimizing low affinity low-molecular-weight fragments into higher affinity lead molecules. Key in this process is the initial identification of fragments that bind to the protein target of interest. Existing computational methods (docking and virtual screening) are optimized for complex drug-like small molecules and do not perform with fragments. However, the availability of structural data for proteins whose bound ligands share substructures can enhance our understanding of fragment binding to facilitate binding predictions.

We propose an unsupervised machine learning approach to automate the discovery of fragment binding preferences. For all protein residues involved in ligand binding, we characterize their local structural microenvironment and annotate them with the ligand fragments they bind. This serves as the knowledge base of protein-fragment interactions. Comparison to the knowledge base enables retrieval of fragments statistically preferred by the microenvironments of a target protein structure, giving insight for drug design. Our approach enables discovery of similar micro-environments across diverse proteins and maximizes structural data usage by merging information across diverse ligands with shared substructures. Results on a dataset of proteins binding a variety ligands show strong ability to rediscover fragments corresponding to the bound ligand, validating the methodology.

METHODS

The foundation of the method is a knowledge base of protein-fragment interactions. From the Protein Data Bank (PDB)[1], we retrieve high-resolution protein structures in complex with a ligand. We characterize the physicochemical environment (microenvironment) around each residue interacting with the ligand using FEATURE[2]. Each residue is also annotated with the ligand substructures (fragments) in close proximity. All ligands from the PDB are divided into overlapping fragments of 3-10 heavy atoms. Protein micro-environments are thereby associated with ligand fragments, forming the knowledge base of protein-fragment interactions (Figure 1).

Given a protein structure with unknown fragment binding preferences, we first identify potential ligand binding pockets. We compute the microenvironments of the residues forming these pockets and compare them to the knowledge base to retrieve similar microenvironments and their bound fragments. There is a core assumption that similar microenvironments share similar fragment binding preferences. We use a hypergeometric distribution and Fisher's method to assess the significance of fragments across microenvironments in spatial proximity. Groups of microenvironments thus have rank-ordered fragment binding preferences (Figure 1). A target protein pocket can yield multiple fragment predictions applicable towards fragment-based drug design.

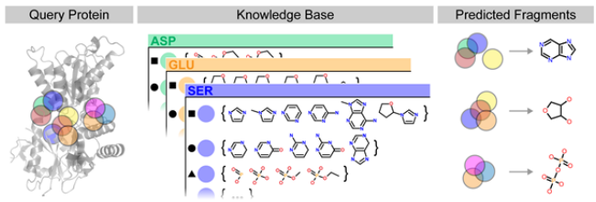

FIGURE 1*. Schematic of method work flow. From a query protein, microenvironments of interest (colored circles) are compared to the knowledge base to retrieve the subset of similar microenvironments. Fragment binding information from similar microenvironments undergoes statistical tests to produce ranked predictions for each group of query microenvironments.*

RESULTS & CONCLUSIONS

We evaluated the performance of our knowledge base and methodology by assessing their ability to rediscover known interactions between proteins and small molecules. From the PDB, we retrieve structures in complex with cofactors adenosine diphosphate (ADP), flavin adenine dinucleotide (FAD), nicotinamide adenine dinucleotide (NAD), and thiamine diphosphate (TPP). These ligands differ in occurrence in the database, have significant flexibility, contain a variety of chemical moieties, and share some moieties in different contexts, testing method robustness under multiple scenarios. We define method performance as the ability to predict fragments corresponding to the ligand bound.

We analyzed 1,423 non-redundant protein chains binding ADP (597), FAD (389), NAD (399), and TPP (38). These proteins chains are scanned by fpocket[3] to determine potential ligand binding sites to assess for fragment binding preferences. Not all portions of a ligand interact with the protein and predicted sites may entirely miss the ligand site. Of the chemical moieties interacting with the analyzed protein sites, we achieve recall of greater than 80% at precision of 70%. Sites not interacting with the ligand were excluded from the analysis. The chemical moieties differ in predictability with thiamine recall the lowest (54%) and phosphate the highest (90%). This behavior is also weakly context specific as the adenine moiety in ADP has 70% recall compared to 80% for FAD and NAD.

Thus, on a validation set of cofactors our unsupervised machine learning approach for predicting fragment binding achieves high performance. However, many of the proteins assessed are in complex with additional ligands that we also successfully predict. These ligands include organo-metallic complexes, enzyme substrates, heme, sugars (ie. glucose), and inhibitors (ie. tricolsan). While performance for these ligands is not yet fully validated, these examples nonetheless demonstrate the ability of our method to produce relevant fragment predictions for a structure of interest.

REFERENCES

1. Berman, H. M. *et al. Nucleic Acids Res* **28**, 235-242 (2000).
2. Halperin, I., Glazer, D. S., Wu, S. & Altman, R. B.. *BMC Genomics* **9** Suppl 2, S2 (2008).
3. Le Guilloux, V., Schmidtke, P. & Tuffery, P.. *BMC Bioinformatics* **10**, 168 (2009).