Predicting Protein-Ligand Binding Specificity Based on Ensemble Clustering

Ziyi Guo  
Dept. of Computer Science and Engineering  
Lehigh University  
Bethlehem, PA, USA  
zig312@lehigh.edu

Brian Y. Chen*  
Dept. of Computer Science and Engineering  
Lehigh University  
Bethlehem, PA, USA  
chen@cse.lehigh.edu

Abstract—Protein structure comparison algorithms can be used to identify distantly related proteins or to categorize differences in binding specificities. When they are presented in different conformations, distantly related proteins can go unrecognized unless flexible representations of whole protein structures are used. Such representations offer a sophisticated description of backbone motion, but they do not incorporate the potential motion of every atom. Thus, existing representations, both rigid and flexible, cannot compensate for atomic motions that can make binding sites with similar binding preferences appear different.

To bridge this gap, this paper presents a tool for comparing protein binding sites despite conformational changes in the binding site. Our method employs ensemble clustering techniques to incorporate the diversity of binding site variations observed in conformational samples of binding site motion. We applied the method on protein conformations of serine proteases and enolase superfamilies. Our results demonstrate that this approach can distinguish proteins with similar binding preferences in the presence of considerable binding site flexibility.

I. INTRODUCTION

Conformational flexibility is significant complication for the accurate the comparison of protein structures. Many algorithms perform efficiently because they apply rigid transformations to superpose atoms from different structures without considering alternative conformations. With this simplifying assumption, existing methods can rapidly align backbone carbon atoms [1]–[7], distance matrices [8], graphical topologies [9]–[11] and geometric surfaces [12]–[15] to detect structural similarities between remote homologs. The rigidity assumption also enables methods to rapidly discover structural variations between closely related proteins with different binding specificities [16], [17]. However, without the simplification of rigidity, the structural comparisons would be more difficult because all protein conformations must be considered.

A recent class of algorithms use rigid secondary structural elements with flexible linkers to represent protein structures via hinges [18], [19], graphs [20], [21], fragments [22] and dynamic programming [23]–[25]. Most approaches are designed to identify remote homologs that could be overlooked from conformations of each protein. But flexible linkers do not describe smaller atomic motions inside binding cavities, and they thus have limited applications to categorize flexible binding sites with subtly different binding preferences.

This paper presents an algorithm for categorizing protein structures based on ligand binding preferences, despite the presence of structural variations. Beginning with conformational samples of several proteins, our method clusters binding sites found in randomly selected samples of each protein. Integrating many randomly generated clusterings yields an average categorization of the structural similarities and variations between the binding sites. This approach reflects all-atom motions in each protein that are represented by the conformational samples, whereas existing methods generally employ artifically rigid and artificially flexible regions. In our results, we demonstrate how this method performs for categorization conformational samples of serine protease and enolase superfamily binding sites.

In the recent years, we have reported new methods that analyze structural flexibility within binding cavities. One method [26] detected clusters of cavity conformations from the same protein. These clusters assisted in predicting influential amino acids that can affect specificity. However, this method did not provide direct comparison between conformations of different proteins. A second method, FAVA [27], defined the three dimensional region that is accessible to most conformational cavities of one protein as the frequent region. Frequent regions could be clustered to produce categorizations of ligand binding cavities, but because the frequent region does not necessarily exist in highly flexible binding sites, we examine methods here based on atomic comparisons.

II. METHODS

Overall, as input, our method begins with conformational samples of one family of protein structures defined by EC classification [28] for comparison. Our method outputs protein clusters that predict ligand binding specificities. First, we designate one protein structure as the template and explain how to compute the template motif: the positions of amino acids that are adjacent to the ligand surface. The motif is considered to be close to the binding cavity and its motion...
may alter the shape of the binding site. Second, we describe how we compute the structural matches of the template motif to identify similar substructures in other protein structures, generating the propagated motifs for each family member.

Given one sampled conformation of each protein, we compute the all-against-all least root mean square distance (LRMSD) similarities between propagated motifs. These similarities create geometric feature vectors that correspond to high dimensional points in the geometric feature space. We continue to build a hierarchical clustering from geometric features and the clustering outputs a cluster label for each family member.

Due to the nondeterministic nature of protein conformation sampling, the clustering could be highly unstable and no single clustering is guaranteed to be reliable across all conformational samples of all protein structures. Therefore, we applied ensemble clustering techniques. Given a set of base clusterings, ensemble methods output a consensus clustering that shares as much information as possible with all base clusterings [29]. Finally, we discuss how to compute such a consensus clustering to predict ligand binding specificity.

### A. Structural motif construction

Formally, within a family of proteins, we select one structure $T$ as the template and refer to its conformational samples as \{$T_1, T_2, ..., T_n$\}. We define the shape of each sampled ligand binding cavity as \{$t_1, t_2, ..., t_n$\} using VASP [17]. Following our earlier work [27], we compute the average intersection volume of each amino acid $r$ between the sampled amino acid $r_i$ of $T_i$ and the sampled ligand binding cavity $t_j$ for all pairs of samples $i$ and $j$. The large average intersection volume indicates $r$ frequently changes the shape of the binding cavity. In this work, we rank all the amino acids by their average intersection volume and return the top $k$ as the template motif $S = \{S_1, S_2, ..., S_k\}$ where $S_x$ is the sequence number for the $x$th top amino acid. The positions of motif $S$ characterize the shape and structure of the binding site. It is noted that our method is independent of intersection volume calculation and adapting other reasonable motif generation methods could also be successful.

### B. Motif propagation

The computed template motif $S$ is matched against a family of protein structures $F = \{f_1, f_2, ..., f_m\}$, yielding a set of matches $M_{S \rightarrow F} = \{M_{S \rightarrow f_1}, M_{S \rightarrow f_2}, ..., M_{S \rightarrow f_m}\}$. In this work, FATCAT [21] is used between the template structure $T$ and each protein structure $f_i$ to identify substructure matches by searching every residue in motif $S$ and returning the matched residue in $f_i$. FATCAT is selected because of the availability and compatibility to flexible structure comparisons. Every substructure match $M_{S \rightarrow f_i}$ is a mapping between $S$ and a substructure of $f_i$, and all the amino acids in the substructure are returned as a propagated motif, $S_{f_i}$. If any arbitrary amino acid $S_i$ in $S$ is aligned to a gap, $S_i$ will be removed from the template motif.

### C. Base clustering generation

To create a base clustering, we take a random sampling of protein conformations $F' = \{f_1, f_2, ..., f_m\}$ as input where $f_{xq}$ indicates the $q$th conformation of the protein $f_x$. All these conformations are superposed onto one selected structure $f_x$ by minimizing the overall root mean square distance (RMSD). We write the pairwise LRMSD between two propagated motifs as $L(S_{f_{j}}, S_{f_{k}})$. The LRMSD is obtained by computing $C_{Q}$ atom RMSD of all amino acids in propagated motifs on $F'$. The geometric feature $g_j$ for protein $f_j$ is a vector defined as $g_j = \{L(S_{f_{1}}, S_{f_{j}}), ..., L(S_{f_{n}}, S_{f_{j}})\}$. The geometric feature space of all-against-all LRMSD alignment within a protein family can be represented by a matrix $G = \{g_1, ..., g_m\}$. Each $g_i$ is a point in the feature space. We hypothesize that proteins with identical binding specificity should be nearby in the feature space and be clustered into the same group.

To test our hypothesis, we use the UPGMA (Unweighted Pair Group Method with Arithmetic mean) [30], an agglomerative hierarchical clustering method with average linkage, to generate a base clustering using geometric features. The UPGMA outputs one base clustering as a label vector $\lambda$ by specifying the number of clusters where the $i$th element $\lambda_i \in \{1, 2, ..., c\}$ indicates the cluster assignment for each feature $g_i$.

### Input: Data set $G = \{g_1, ..., g_m\}$;
Base UPGMA clusters $\Omega^{(q)}$, $q = 1, ..., r$;
The consensus UPGMA cluster $\Omega$;

#### Process:
1. For $q = 1, ..., r$:
2. $\lambda^{(q)} = \Omega^{(q)}(G)$;
3. Form an $m \times m$ base similarity matrix $M^{(q)}$ from $\lambda^{(q)}$ using Equation (1);
4. End
5. $M = \frac{1}{r} \sum_{q=1}^{r} M^{(q)}$;
6. $\lambda^* = \Omega(M)$;

#### Output: Ensemble clustering vector $\lambda^*$.

Fig. 1. CSPA Ensemble Clustering Algorithm.

### Serine Protease Superfamily:
- 

### Enolase Superfamily:
- 

### Mucodlactonizing Enzyme
- 2pgw
D. Ensemble clustering

In this step, we have \( r \) base clustering vectors \( \{\lambda(1), \lambda(2), \ldots, \lambda(r)\} \) using conformation sampling with replacement. In order to ensemble all the base clusterings, we need a combination function \( \Gamma \) to create a consensus clustering \( \lambda^* = \Gamma(\{\lambda(1), \lambda(2), \ldots, \lambda(r)\}) \). Given \( m \) protein structures and \( n \) conformations of each structure, the number of all possible samplings is \( n^m \), and the exponential size of combination is impractical even for very small \( n \) and \( m \). Therefore, the brute force search over all possible samplings is infeasible and a heuristic strategy is needed.

Here, we adopt a cluster-based similarity partitioning algorithm (CSPA) to compute a consensus clustering. Essentially, if two objects are in the same cluster, they are considered to be fully similar, and if not they are fully dissimilar. To achieve this, we convert a base clustering vector \( \lambda^{(q)} \) of size \( m \) to an \( m \times m \) base similarity matrix \( M^{(q)} \) by:

\[
M^{(q)}_{(i,j)} = \begin{cases} 
0 & \text{if } \lambda_i^{(q)} = \lambda_j^{(q)} \\
1 & \text{otherwise} 
\end{cases} 
\]

we average all the base similarity matrices, yielding an averaged similarity matrix \( M \). Here, the less \( M_{(i,j)} \) is, the more possibility that the \( i \)th object and the \( j \)th object will be grouped into the same cluster. Finally, we form a consensus UPGMA clustering based on the averaged similarity matrix. The general process of CSPA is illustrated in Figure 1. For more details about ensemble clusterings and CSPA, see [29], [31].

E. Data set

We test our method on two protein superfamilies, the serine protease and the enolase superfamily. In the serine protease, the trypsin, chymotrypsin and elastase subfamilies were selected. In the enolase superfamily, the enolase, mandelate racemase and muconate lactonizing enzyme subfamilies were selected.

The serine proteases hydrolyze peptide bonds by recognizing a set of adjacent amino acids with specificity subsites that are numbered \( S_4, S_3, S_2, S_1, S'_1, S'_2, S'_3, S'_4 \). Each subsite has binding preferences on one amino acid before or after the \( S_1 - S'_1 \) hydrolyzed bond. In this work, we focus on three different binding specificities of the \( S_1 \) subsite: positively charged amino acid [32] for trypsins, large and hydrophobic amino acid [33] for chymotrypsins and small hydrophobics [34] for elastases.

The enolase superfamily proteins catalyze reactions by abstracting a proton from a carbon adjacent to a carboxylic acid [35] near the C-terminal domain of beta sheets of the conserved TIM-barrel structures. In this work, we study three different catalysts. The enolase subfamily converts 2-phosphoglycerate (2-PG) to phosphoenolpyruvate (PEP) [36], the mandelate racemases convert between (S)-mandelate and (R)-mandelate [37] and the muconate-lactonizing enzymes convert lignin-derived aromatics, catechol and protocatechuate to citric acid cycle intermediates [35].

**Protein Selection.** 676 serine protease structures and 66 enolase superfamily structures were selected from the Protein Data Bank (PDB) [38] on 06.21.2011. Protein structures with mutation, disordered regions or closed regions were removed. Then, structures were filtered to keep less than 90% pairwise sequence identity with preference for maintaining structures with literature description. Several structures (8gch, 1aks etc.) were removed because of technical problems of protein simulation. On the remaining 12 serine proteases and 7 enolases, non protein atoms, such as ions and water, and hydrogens were removed. All the structures are shown in Figure 2 by their PDB code and are classified into subfamilies by their binding specificities.

When generating one base clustering from one sampling of protein conformations, we superposed all the sampled conformations using Ska, an whole structure alignment algorithm [39]. We aligned all serine protease conformations onto 8gch and all enolase superfamily conformations onto 1mdr. These two structures were selected because of the existence of ligand bound. All the structures in the same superfamily have identical protein folds and the choice of alignment methods is of little difference [17].

F. Protein structure simulation

The conformational samples of each protein structure were simulated using GROMACS 4.5.4 [41]. The input structure was centered inside a cubic waterbox using a 3-point solvent model SPC/E [42]. The waterbox was set so that there is at least 10 Å between the protein and the nearest part of the box. Charge balanced sodium and potassium were then added at a low concentration (< 0.1% salinity). Steepest descent was used to minimize energy on the entire simulation system. Isothermal-Isobaric (NPT) equilibration was performed in four 250 picoseconds steps for temperature and pressure equilibration before the primary simulation. Over the four 250 picosecond minimization period, at 1000 kJ/(mol * nm), each equilibration step reduced the position restraint force by 250 kJ/(mol * nm). Backbone positions constraints were released during the NPT simulation and system energies were computed in the beginning of the equilibration phase. Temperature was set to 300 Kelvin and pressure was set to 1 bar. Temperature coupling was computed using Nosé-Hoover thermostat [42] and pressure coupling was computed using the Parrinello-Rahman algorithm [43], [44]. The simulation used

<table>
<thead>
<tr>
<th>PDB</th>
<th>Motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a4j</td>
<td>S190 G183 S195 V213 W215 G216 K224 P225</td>
</tr>
<tr>
<td>1ebh</td>
<td>D246 C247 Q295 D320 K345 H373 R374 K396</td>
</tr>
</tbody>
</table>
P-LINCS [45] to update bonds and used particle mesh Ewald summation (PME) [41] to calculate electrostatic interaction energies. The primary MD simulation was started using the atomic positions and velocities of the final equilibrium state. The simulation was maintained for 100 nanoseconds with 1 femtosecond timesteps on multiple 16 core nodes of the Lehigh corona server where OpenMPI was used for parallel communications. The trajectory file was convert to the PDB format with only atomic positions.

III. EXPERIMENTAL RESULTS

First, we show that conformational binding cavities vary considerably over all protein simulations, and these variations can be taken as evidence that weakens the rigidity assumption for protein structure comparisons. Then, we demonstrate the template motif on two families and continue to visualize the propagated motif of other proteins. Finally, we show the ensemble clustering to predict ligand binding specificity and compare with some previous works.

A. Considerable variation of conformational binding cavities

Figure 3 illustrates variations of protein binding cavity volumes over all conformations in our whole data set and we observe all proteins cavities varied considerably. Specifically, trypsin cavity volumes ranged from 249 Å$^3$ to 693 Å$^3$, chymotrypsin cavity volumes ranged from 127 Å$^3$ to 553 Å$^3$ and elastase cavity volumes ranged from 277 Å$^3$ to 569 Å$^3$. The significant volume variations can also be detected in the enolase superfamily. Enolase cavity volumes ranged between 90 Å$^3$ to 508 Å$^3$, mandelate racemase volumes ranged between 225 Å$^3$ to 673 Å$^3$ and mucionate lactonizing enzyme volumes ranged between 90 Å$^3$ to 344 Å$^3$.

All these observations demonstrate considerable variations of binding cavities of the same protein. The cavity variations create errors for flexible structure comparisons, preventing accurate predictions of ligand binding specificity when protein conformational samples are considered [27]. The protein binding cavity varied because of the motion of adjacent amino acids, leading to the motivation to identify structural motifs for binding specificity prediction.

B. Motif definition and propagation

We selected 1a0j as the template structure for the serine protease and 1ebh as the template structure for the enolase superfamily. We ranked all the amino acids of two protein structures and added the top 8 residues into the template motif. In table I, we show all amino acids of the template motif. Figure 4 illustrates 3D structure of the motif in one conformational sample of 1a0j and 1ebh. We observe that both motifs are close to the binding cavity, and their motions may enlarge, shrink or even separate binding cavities.

Three amino acids, \{C$^{247}$, R$^{374}$, K$^{396}$\}, in the template motif of enolases were removed during motif propagation because they were aligned to the gap. Figure 5 shows the superposition of the propagated motifs of all proteins in our data set. The motifs of proteins with the same binding preference tend to form closely-located substructure clusters, leading to cluster analysis to predict ligand binding specificity.

C. Protein ensemble clustering

Figure 6B demonstrates the ensemble UPGMA clustering of propagated motifs on serine protease structures. Proteins in the same subfamily are correctly clustered into the same group. Figure 6C demonstrates the UPGMA clustering of frequent regions using FAVA. We observe that the only chymotrypsin protein, 1ex3, was misclassified into the trypsin cluster and two elastases were separated into different clusters. Moreover, Figure 6B exhibits a greater similarity between trypsins than Figure 6C. This indicates that structural motifs may be better markers to distinguish proteins with different binding preferences.

Figure 7B shows the ensemble clustering of propagated motifs on enolase superfamily structures. Three subfamilies are all correctly clustered by their binding specificities. Figure 7C shows the clustering using FAVA. We can see that two mandelate racemases were separated and one of them, 1mdr, was misclassified into the enolase cluster. Similarly, greater similarities between the enolases and between mandelate racemases was detected using the ensemble clustering.
IV. CONCLUSION

We have presented a computational method to compare conformational protein structures based on the ensemble clustering. Unlike existing methods with rigidity assumption of protein structures or secondary structure elements, our method extracts propagated motifs of protein conformations to characterize the motion of protein binding sites. This capacity enables a novel representation of molecular flexibility of binding cavities using conformational samples.

We applied our method on conformational samples of sequentially nonredundant structures of two protein superfamilies: the serine protease and the enolase superfamily. The proteins in both superfamilies revealed considerable structural variations in binding cavities. Despite these flexibilities, our method correctly classified all protein structures of both superfamilies according to their substrate binding preferences. This result indicates that atomic comparisons of highly similar proteins can exhibit subtle differences that affect specificities when conformationally structural flexibilities are considered.

Our method has great application potentials for comparisons of proteins with identical folds but different ligand binding preferences. In such cases, our method generates propagated motifs to represent shape of protein binding cavities, pointing to the local structure that is relevant to substrate binding. Using ensemble clustering techniques, our method mitigates the prediction errors from conformational flexibilities. These
capacities provide an important tool for structure-based function annotation of molecular design.

ACKNOWLEDGEMENT

This work was supported in part by National Science Foundation Grant 1320137 to Brian Chen and Katya Scheinberg.


