A Map of Binding Cavity Conformations Reveals Differences in Binding Specificity

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

Abstract—Protein structure comparison algorithms are useful for predicting aspects of protein function. Some algorithms identify remote homologs, while others distinguish closely related proteins that prefer different substrates. Most of these methods assume that proteins are rigid in order to perform comparisons more rapidly, while others compensate for flexibility by representing proteins as a connected group of rigid components. To consider the motion of individual atoms, this paper presents a method for generating a map of binding cavity conformations based on conformational snapshots. We use clusters of protein conformations to distinguish proteins that have different binding preferences. Our results, on the serine proteases and enolase superfamilies show that, despite structural flexibility in binding sites, our methods correctly classify proteins with different binding specificities both qualitatively and quantitatively.

I. INTRODUCTION

Conformational flexibility is a universal challenge in the comparison of protein structures because flexibility interferes with typical notions of geometric similarity. Most comparison algorithms model proteins as rigid objects. This simplifying assumption facilitates rapid comparison and it is most compatible with the available data, because alternative conformations are infrequently available from experimental sources. Rigid representations of protein structure use carbon alpha coordinates [1]–[7], distance matrices [8], graphs [9]–[11] and geometric surfaces [12]–[15] to detect similarities between remote homologs or to identify proteins with different binding preferences [16]–[20]. A second class of comparison methods use hinges [21], [22], graphs [23], [24], fragments [25] or dynamic programming [26]–[28] to represent proteins as rigid structural elements connected by flexible regions. Since these representations do not compensate for structural motion within binding sites, accurate classification of binding cavities, when they appear with different conformations, can be a challenge.

The specific problem we are addressing is to distinguish binding cavities with different binding specificities in the presence of conformational flexibility. This paper presents an algorithm for generating a map of binding cavity conformations from multiple proteins. Our method happens to use molecular dynamics trajectories to generate conformations, but it is not simply a technique for analyzing molecular dynamics trajectories. Other techniques for generating conformations, such as Monte Carlo sampling methods, would be equally effective. Our method works by first selecting amino acids that describe the binding site, called a motif. Cognate amino acids are identified in all proteins being compared in a process we call motif propagation. Finally, the coordinates of each amino acid, represented by its alpha carbon location, are identified in every conformation of every protein. These coordinates are collected as a data points in a high dimensional space. Finally, we use Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) to project the data points into a $3D$ space. We hypothesize that binding sites with identical binding preferences will cluster together.

This approach contrasts from our earlier studies in representing all-atom motion at binding cavities. Our initial efforts combined multiple conformational snapshots into a single aggregate representation of the binding cavity, using conserved $3D$ regions [29], cavity clusters [30] or prediction ensembles [31]. While these representations reflect detailed conformational data, our comparisons of aggregate representations only produced single dimensional measures of similarity rather than the nuanced picture of similarity at different conformations. The mapping technique proposed here avoids generalizations produced by aggregate representations.

Our approach is inspired in part by existing methods that define protein structure space as either a discrete or continuous geometric space of protein folds (e.g. [8], [32]–[34]). Protein structure space leverages structure comparisons to provide a comprehensive view on how protein structures are distributed, which is significantly different from hierarchical classification systems such as Structural Classification of Proteins (SCOP) [35] and CATH [36]. One way to approach protein structure space understanding is to represent the space in lower dimensional space, such as three dimensional maps. Several efforts [33], [37]–[42] have been reported to construct maps of protein structure space. These methods develop clever techniques for computing structure similarities between all pairs of conformations, and then use dimension reduction methods to create lower-dimensional embeddings for visualization. Here, we adapt these concepts to the comparison of binding sites with varying binding preferences.

To our knowledge, our method is the first effort for analyzing maps of binding cavity conformations to classify...
proteins with different specificities. We tested our method on sequentially nonredundant protein structures of serine proteases and enolases. In both datasets, we observe that cavities with the same binding preference form closely-located clusters in the conformation map produced by our software. The quantitative clustering evaluations show that our method can distinguish proteins with different binding specificities despite considerable variations of their binding cavities.

II. METHODS

Overall, our method accepts conformational samples of one family of protein structures as input. These proteins exhibit identical folds but reveal different binding specificities. Our method outputs a conformation space map that models conformational flexibilities of binding cavities. First, we describe how we define the template motif which is a collection of residue positions of the template structure that are adjacent to the ligand molecule, in the selected template structure. Each amino acid in the template motif is close to the binding cavity and its motion may affect the shape of the binding site. Second, we explain how we perform whole structure alignment to identify analogous substructures, thus propagating structural motifs to other proteins. We extract three dimensional Carbon alpha coordinates for each member in the structural motif for all protein conformations. Therefore, the motif of each conformation can be characterized with a feature vector. Since motifs are propagated by detecting substructural matches, some features could be highly similar. These features will increase feature space dimensionality but are not necessarily discriminative for different binding specificities, leading to perform dimension reduction in step three. Here, we select two effective reduction methods, Non-negative Matrix Factorization (NMF) and Principal Component Analysis (PCA).

Each binding cavity conformation can then be represented as a data point in the reduced feature space. It is hypothesized that conformations of proteins with identical binding specificity should be close to each other and be grouped into the same cluster. Finally, we discuss how we perform data clustering to test our hypothesis.

A. Template motif generation

Formally, from one family of protein structures, we select one protein $T$ as the template structure and its conformational samples are referred to as $\{T_1, T_2, ..., T_N\}$. The binding cavity conformations, $\{t_1, t_2, ..., t_m\}$, can be generated using VASP [16]. For each amino acid $i$ of $T$, we compute the median intersection volume between $i$ of the conformation $T_i$ and the binding cavity conformation $t_j$ for all pairs of $i$ and $j$. The non trivial average intersection volume indicates that $r$ frequently overlaps with the binding cavity so it could change the shape of the binding cavity substantially. We continue to rank all amino acids by their average intersection volume, and add top $k$ into the template motif $\mathcal{S} = \{S_1, S_2, ..., S_k\}$. The motif size is defined as the number of selected amino acids.

B. Motif propagation

To identify similar motifs in proteins that are not template structures, $T$ is structurally aligned against a family of protein structures $F = \{f_1, f_2, ..., f_M\}$. Here, we run FATCAT [24] between $T$ and each protein structure to find substructural match $M_{S \rightarrow f_j}$ by searching every residue in the template and returning the matched residue in $f_j$. FATCAT is used because it is compatible and available to flexible structure comparisons. $M_{S \rightarrow f_j}$ defines a structural match between the template motif and a substructure in $f_j$, and all the matched residues in this substructure is called the propagated motif. If any residue in the template motif is matched to a gap, it will be removed from the motif. It is noted that our method is independent of FATCAT alignment, other substructure matching algorithms, such as LabelHash [43] and Match Augmentation [44], can also work.

C. Dimension Reduction

Given propagated motifs, one binding cavity conformation can be characterized as a geometric feature vector where each value is the $x$ or $y$ or $z$ direction coordinate of $C_{\alpha}$ atom. All feature vectors will be normalized so that each data point has unit norm. The feature matrix $X = \{x_1, ..., x_n\} \in \mathbb{R}^{m \times n}$ represents geometric features of all conformation samples of all proteins where $n$ is the total conformation number and $m$ is the feature dimensionality, and the matrix will be taken as input for dimension reduction.

Non-negative matrix factorization (NMF) [45] is a matrix decomposition algorithm for parts-based data representation of matrices with non-negative elements. Given input matrix $X$, NMF aims to find two non-negative components $W \in \mathbb{R}^{m \times r}$ and $H \in \mathbb{R}^{r \times n}$ to minimize the objective function where $r$ is the reduced feature dimensionality:

$$\min_{W,H} F = \|X - WH\|^2$$

s.t. $W_{ij} \geq 0, H_{ij} \geq 0$ (1)

The objective is convex with respect to either $W$ or $H$, but not convex in both together so that the global optimal is difficult to find. Starting from random initialization of $W$ and $H$, Lee

<table>
<thead>
<tr>
<th>Serine Protease Superfamily:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chymotrypsins: 1ex3</td>
</tr>
<tr>
<td>Elastases: 1b0c, 1elt</td>
</tr>
<tr>
<td>Trypsins: 1a0j, 1ane, 1aq7, 1bzx, 1fn8, 1h4w, 1trn, 2eek, 2f91</td>
</tr>
<tr>
<td>Enolase Superfamily:</td>
</tr>
<tr>
<td>Enolases: T68h, Tiyx, Tte6, 3otr</td>
</tr>
<tr>
<td>Mandelate Racemase: 1mdr, 2ox4</td>
</tr>
<tr>
<td>Muconate Lactonizing Enzyme: 2pgw</td>
</tr>
</tbody>
</table>

Fig. 1. PDB codes used in the data set.
and Seung [46] presented an algorithm to iteratively update $W$ and $H$ and a local minimum can be guaranteed:

$$W_{ij} = W_{ij} \frac{(XH^T)_{ij}}{(WHHT)_{ij}}$$

$$H_{ij} = H_{ij} \frac{(W^TX)_{ij}}{(WTH)_{ij}}$$

(2)

Usually, we have $r \ll m$ and $r \ll n$. Thus, NMF essentially try to discover latent structures using very few dimensions as a compressed representation. If there exists negative elements in the matrix, we translate the matrix so that the non-negative constraint is guaranteed.

Principal Component Analysis (PCA) [47] is one of the most popular dimension reduction methods. PCA orthogonally project a set of data points onto a lower $r$ dimensional principal subspace such that variances between projected data are maximized. The projection vectors can be computed as a set of eigenvectors with top $r$ largest eigenvalues.

D. Cluster analysis

We perform the canonical K-means clustering to identify data clusters. The performance is evaluated using clustering accuracy (AC).

Given the predicted cluster label $l_i$ and the ground truth $g_i$, AC is defined as:

$$AC = \frac{\sum_{i=1}^{n} \delta(g_i, \text{map}(l_i))}{n}$$

(3)

Where $\delta(\cdot)$ is the delta function that equals to one for identical comparison and equals to zero otherwise and map($\cdot$) is a mapping function that permutes the predicted label set to match the ground truth set as much as possible. This can be done using the Hungarian method [48].

III. Dataset

A. Protein superfamily selection

We tested our method on sequentially nonredundant structures of two protein superfamilies: the serine protease and the enolase superfamily. In serine proteases, trypsin, chymotrypsin and elastase superfamilies were selected. In the enolase superfamily, the enolase, mandelate racemase and muconate lactonizing superfamilies were selected.

The serine protease is a family of enzyme proteins that selectively cleaves peptide bonds where serine functions as the nucleophilic amino acid at the binding site. The preferences for hydrolyzing a specific scissile bond can be achieved by recognizing amino acids on both sides of the bond. The most well-known residue is the $P1$ just before the bond. The $S1$ binding pocket, which recognizes $P1$, exhibits three different binding specificities: positively charged amino acids for trypsins [49], large hydrophobic amino acids for chymotrypsins [50] and small amino acids for elastases [51].

The enolase superfamily proteins catalyze biochemical reactions with an abstraction of a proton from a carbon that is adjacent to a carboxylic acid and a requirement of a divalent metal ion [52]. Here we focus on the specificities of three catalysts. The enolase subfamily converts 2-phosphoglycerate (2-PG) to phosphoenolpyruvate (PEP) [53], the mandelate racemases convert between (S)-mandelate and (R)-mandelate [54] and the muconate-lactonizing enzymes convert lignin-derived aromatics, catechol and protocatechuic to citric acid cycle intermediates [52].

B. Protein structure selection

We downloaded all protein structures of the serine protease and the enolase superfamily from Protein Data Bank [55]. We removed all the structures with mutation, disordered regions or closed regions. We further kept one structure from any pair of structures with more than 90% sequence similarity where structures associated with publications were preferred. Few structures, such as 8gch and 1aks, were removed because of technical issues of MD simulation. In the end, we have 12 serine proteases and 7 enolase structures, and they are shown in Figure 1 with classification into subfamilies by their binding specificities.

We superposed all the conformational samples using ska [56], a whole structure alignment tool. We superposed all conformation structures of serine proteases onto 8gch and all conformation structures of the enolases onto 1mdr. These two structures were selected because of ligand bound existence.

C. Protein structure simulation

Conformational samples of all protein structures were simulated using GROMACS 4.5.4 [57]. The input protein was centered inside a cubic waterbox using a 3-point solvent model SPC/E [58] with at least 10 Å between the structure and the nearest part of the waterbox. Charge balanced sodium and potassium were added with lower than 0.1% salinity. Isothermal-Isobaric (NPT) equilibration in four 250 picoseconds steps was run for temperature and pressure equilibration. Each equilibration step reduced the position restraint force by 250 kJ/(mol*nm) where backbone positions constraints were released and system energies were computed. Temperature was set to 300 Kelvin and pressure was set to 1 bar. Temperature coupling was computed using Nosé-Hoover thermostat [58] and pressure coupling was computed using the Parrinello-Rahman algorithm [59], [60]. The simulation update bonds using P-LINCS [61] and calculate electrostatic energies using particle mesh Ewald summation (PME) [57]. The primary MD simulation was performed for 100 nanoseconds with 1 femtosecond timesteps on multiple 16 core nodes of the Lehigh corona server. The trajectory file was convert to the PDB format with only atomic positions. For each protein structure, 600 samples were selected at uniform intervals.

IV. Results

In this section, we first demonstrate considerable variations of binding cavity volumes in our data set, and these variations could be sources of errors that weaken rigidity assumption for protein structure comparisons. Second, we show 3D structures of template motifs and propagated motifs where co-located
substructure clusters are observed that reveal specificities. Finally, we evaluate clustering performance for predicting binding specificity.

A. Binding cavity varies considerably

Considerable variations of protein cavity volumes can be detected over all conformations in our data set as shown in Figure 2. To be specific, in serine proteases, trypsin volumes ranged from 249 Å³ to 693 Å³, chymotrypsin cavity volumes ranged from 127 Å³ to 553 Å³ and elastase cavity volumes ranged from 277 Å³ to 569 Å³. Similarly, in the enolase superfamily, enolase cavity volumes ranged between 90 Å³ to 508 Å³, mandelate racemase cavity volumes ranged between 225 Å³ to 673 Å³ and muconate lactonizing enzyme cavity volumes ranged between 90 Å³ to 344 Å³. All these observations reveal structural variations of binding cavities in the same protein. The cavity variations create errors for flexible binding cavity comparison, preventing accurate specificity prediction when protein conformational samples are used [29]. Protein cavities varied because of motions of adjacent amino acids, and thus we identify these amino acids for representing binding cavities.

B. Motif definition and propagation

We selected 1a0j as the template structure for serine proteases and selected 1ebh as the template structure for the enolases. All the proteins in the same superfamily have identical folds and the choice of the template structure has little influence in generating structural motifs. Figure 3 illustrates the 3D structure of template motifs that have 6 residues of 1a0j and 1ebh. We observe that both two template motifs are close to the binding ligand, and their variations may enlarge, shrink or even separate binding cavities.

Figure 4 illustrates superposition of propagated motifs. The superposition exhibits geometrical diversities and motifs structures in proteins with identical binding specificity tend to form closely-located substructure clusters. We hypothesize that it is the structural differences within the selected amino acids in the structural motif that cause proteins in different subfamilies to exhibit different binding specificities. We apply data clustering on binding cavity conformation map to categorize different clusters to predict binding specificity.
C. Clustering evaluation

The map of binding cavity conformation on the enolases is illustrated in Figure 5. We observe that in both NMF and PCA map space, conformations of proteins with the same binding specificity are represented by spatially adjacent points. A similar map can also be found on serine proteases. Such representations reveal a high level organization for binding cavity conformation classification. Adjacent points form specificity-sensitive clusters that can be further evaluated when compared with ground truth EC numbers.

We conducted evaluations with different size of structural motifs. Figure 6 and Figure 7 report clustering results in the original feature space (K-means), PCA reduced space (PCA+K-means) and NMF reduced space (NMF+K-means) on our data set. Clustering always take 3 as the number of clusters because both serine proteases and the enolases exhibit 3 subfamilies. For each motif size $k$, 100 clustering runs were conducted and 100 NMF runs were conducted for every clustering since both K-means and NMF are dependent of data initialization. The average performance is reported for K-means and PCA+K-means, and NMF+K-means performance is shown in boxplot.

These two figures reveal several insights. First, in the original feature space, performance increases as more amino acids are added into structural motifs but suddenly decreases when the size of motif is larger than a threshold. The threshold is about 13 on both superfamilies, indicating the maximal number of amino acids that are relevant to binding. This means that, if the motif size is too small, the binding cavity will be under-represented because some other influential amino acids are not included. If the motif size is too large, the binding cavity will be over-represented with systematic noises because these exist amino acids that are irrelevant to binding. Second, in most cases (except when $k$ ranged between 8 to 12 on serine proteases), PCA+K-means achieve comparable or even better performance to K-means in the original space. This suggests PCA extracts most data variances that are sufficient enough to distinguish protein conformations with different specificities. Third, in average, NMF performs not as well as the other two methods. This is because NMF is largely affected by data initialization. However, if we only consider the best result, NMF+K-means achieves better performances when $k$ is larger than 12 on serine proteases and for almost all $k$ values on enolases.

Overall, our conformation space map reveals high-level representations of binding cavity motions. The clustering results show that our method is able to correctly classify similar proteins with different binding specificities. Therefore, our method could be a robust tool for protein structure comparisons, despite great flexibility in the binding cavity.
enolases.

Fig. 7. Clustering accuracy with respect to the size of structural motif on
essentially different from existing works. First, without rigidity
our conformation space map focuses on examining proteins
model conformational flexibility of binding cavities. Seco-
that change conformations or proteins with different folds.

dation Grant 1320137 to Brian Chen and Katya Scheinberg.
protein cavities, pointing to individual residues that affect
structures of serine proteases and the enolases. Protein struc-
tural motifs to represent conformational flexibilities of
specificities, which is not restricted by hierarchical cate-
distribution of protein cavity conformations with different
binding cavity conformation map. The clustering evalu-
in both superfamilies exhibited highly flexible binding sit-
We tested our method on sequentially nonredundant struc-
tures of serine proteases and the enolases. Protein structures
in superfamilies exhibited highly flexible binding sites.
Despite these structural variations, proteins with identical
binding specificity are represented as adjacent points in our
binding cavity conformation map. The clustering evaluations
show that our method predicts binding specificity with high
accuracy.

Applications of our method exist in comparisons of similar
proteins with different binding preferences. In such cases,
our conformation space map gives a comprehensive visual
distribution of protein cavity conformations with different
specificities, which is not restricted by hierarchical catego-
rizations of the EC number. Moreover, our method generates
structural motifs to represent conformational flexibilities of
protein cavities, pointing to individual residues that affect
binding. These characteristics can be useful for structure-based
function annotation of molecular design.

ACKNOWLEDGEMENT
This work was supported in part by National Science Foun-
dation Grant 1320137 to Brian Chen and Katya Scheinberg.

REFERENCES
program for protein structure comparison,” Computer methods for
macromolecular sequence analysis, 1996.
structural motifs in biological macromolecules by computer vision
techniques,” Proceedings of the National Academy of Sciences, vol. 88,
incremental combinatorial extension (ce) of the optimal path.” Protein
electrostatics of macromolecular structures and sequences.” Methods in
terns: new examples of convergent evolution,” Journal of molecular
[6] B. Y. Chen, V. Y. Fofanov, D. H. Bryant, B. D. Dodson, D. M. Kris-
tensen, A. M. Lisewski, M. Kimmel, O. Lichtarge, and L. E. Kavraki,
“The Wash pipeline for protein function prediction and an algorithm
for the geometric refinement of 3d motifs.” Journal of Computational
clustering of residue position subsets predicts inhibitor affinity across the
human kinome,” PLoS computational biology, vol. 9, no. 6, p. e1003087,
2013.
description of protein structures and its application in predicting ligand
structure comparison,” Current opinion in structural biology, vol. 6,
of protein surfaces using a genetic algorithm,” Journal of Computer-
[12] T. A. Binkowski, S. Naghibzadeh, and J. Liang, “Cas:tp atlas of
surface topography of proteins,” Nucleic Acids Research, vol. 31,
[13] T. A. Binkowski, L. Adamian, and J. Liang, “Inferring functional rela-
tionships of proteins from local sequence and spatial surface patterns,”
shape matching and local spatial alignments of ligand binding sites,”
sites on the molecular surface of proteins,” Protein Science, vol. 14,
properties yields insights into protein-ligand binding specificity,” PLoS
computational biology, vol. 6, no. 8, p. e1000881, 2010.
binding pockets from order-independent surface alignment: a study of
metalloendopeptidase and nad binding proteins,” Journal of molecular
[18] B. Chen and S. Bandyopadhyay, “VASP: A Volumetric Analysis and
Statistical Model for Predicting Steric Influences on Protein-Ligand
Binding Specificity,” in Proceedings of 2011 IEEE International Con-
ference on Bioinformatics and Biomedicine (BIBM), 2011 IEEE Interna-
volume in ligand binding cavities,” in Bioinformatics and Biomedicine
Workshops (BIBM), 2011 IEEE International Conference on. IEEE,
[21] K. Gunasekaran and R. Nussinov, “How different are structurally flexible
and rigid binding sites? sequence and structural features discriminating
proteins that do and do not undergo conformational change upon ligand
2007.
[22] M. Shatsky, R. Nussinov, and H. J. Wolfson, “Flexprot: alignment of
flexible protein structures without a predefinition of hinge regions,”
similar protein binding sites by local structural alignment.” Bioinforma-


