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 categorizing 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

* Corresponding author.

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

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 that 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 xth 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.

Input: Data set $\mathbf{G} = \{\mathbf{g}_1, ..., \mathbf{g}_m\}$;
Base UPGMA clusters $\Omega^{(q)}$, $q = 1, ..., r$;
The consensus UPGMA cluster Ω;
Process:
1. **For** $q = 1, ..., r$:
2. $\lambda^{(q)} = \Omega^{(q)}(\mathbf{G})$;
3. Form an $m \times m$ base similarity matrix $\mathbf{M}^{(q)}$ from $\lambda^{(q)}$ using Equation (1);
4. **End**
5. $\mathbf{M} = \frac{1}{r}\sum_{q=1}^{r} \mathbf{M}^{(q)}$;
6. $\lambda^* = \Omega(\mathbf{M})$;
Output: Ensemble clustering vector λ^*.

Fig. 1. **CSPA Ensemble Clustering Algorithm.**

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 $\mathbf{M}_{S\to F} = \{M_{S\to f_1}, M_{S\to f_2}, ..., M_{S\to f_m}\}$. 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 $M_{S\to f_i}$ 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 $F' = \{f_{1_i}, f_{2_{i'}}, ..., f_{m_{i''}}\}$ as input where f_{x_y} indicates the yth conformation of the structure 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_α atom RMSD of all amino acids in propagated motifs on F'. The geometric feature $\mathbf{g}_j$ for protein f_j is a vector defined as $\mathbf{g}_j = \{L(S_{f_j}, S_{f_1}), ..., L(S_{f_j}, S_{f_m})\}$. The geometric feature space of all-against-all LRMSD alignment within a protein family can be represented by a matrix $\mathbf{G} = \{\mathbf{g}_1, ..., \mathbf{g}_m\}$. Each $\mathbf{g}_j$ 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 (Unweighed 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 λ by specifying the number of clusters where the ith element $\lambda_i \in \{1, 2..., c\}$ indicates the cluster assignment for each feature $\mathbf{g}_i$.

Serine Protease Superfamily:
Chymotrypsins: 1ex3
Elastases: 1b0e, 1elt
Trypsins: 1a0j, 1ane, 1aq7, 1bzx, 1fn8, 1h4w, 1trn, 2eek, 2f91
Enolase Superfamily:
Enolases: 1ebh, 1iyx, 1te6, 3otr
Mandelate Racemase: 1mdr, 2ox4
Muconate Lactonizing Enzyme: 2pgw

Fig. 2. **PDB codes used in the data set.**

D. Ensemble clustering

In this step, we have r base clustering vectors $\{\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(r)}\}$ using conformation sampling with replacement. In order to ensemble all the base clusterings, we need a combination function Γ to create a consensus clustering $\lambda^* = \Gamma(\{\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(r)}\})$. 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 $\mathbf{M}^{(q)}$ by:

$$\mathbf{M}^{(q)}_{(i,j)} = \begin{cases} 0 & \text{if } \lambda^{(q)}_i = \lambda^{(q)}_j \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

we average all the base similarity matrices, yielding an averaged similarity matrix $\mathbf{M}$. Here, the less $\mathbf{M}_{(i,j)}$ is, the more possibility that the ith object and the jth 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 I
THE TEMPLATE MOTIF

PDB	Motifs
1a0j	S190 G193 S195 V213 W215 G216 K224 P225
1ebh	D246 C247 Q295 D320 K345 H373 R374 K396

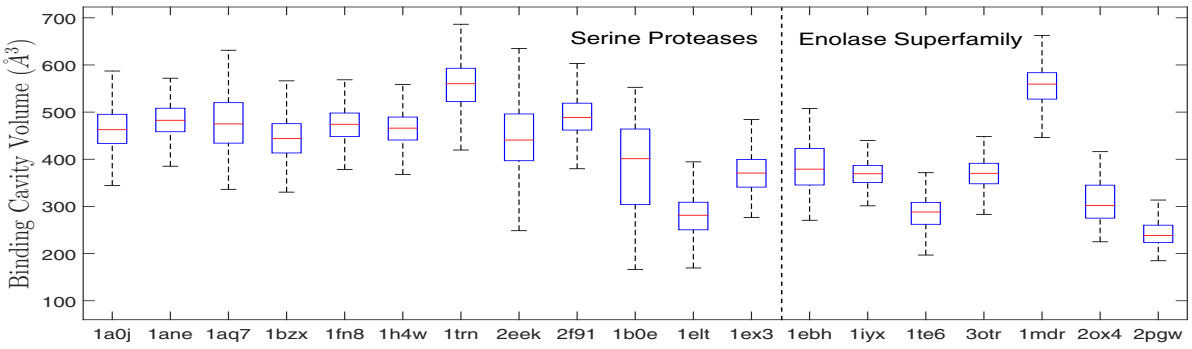

Fig. 3. Aggregate variations in sampled cavity volume in our whole data set. All protein cavity samples varied considerably.

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. 600 samples were selected of each protein structure at uniform intervals.

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 $\AA^3$ to 693 $\AA^3$, chymotrypsin cavity volumes ranged from 127 $\AA^3$ to 553 $\AA^3$ and elastase cavity volumes ranged from 277 $\AA^3$ to 569 $\AA^3$. The significant volume variations can also be detected in the enolase superfamily. Enolase cavity volumes ranged between 90 $\AA^3$ to 508 $\AA^3$, mandelate racemase volumes ranged between 225 $\AA^3$ to 673 $\AA^3$ and muconate lactonizing enzyme volumes ranged between 90 $\AA^3$ to 344 $\AA^3$.

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, $\{C247, R374, K396\}$, 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.

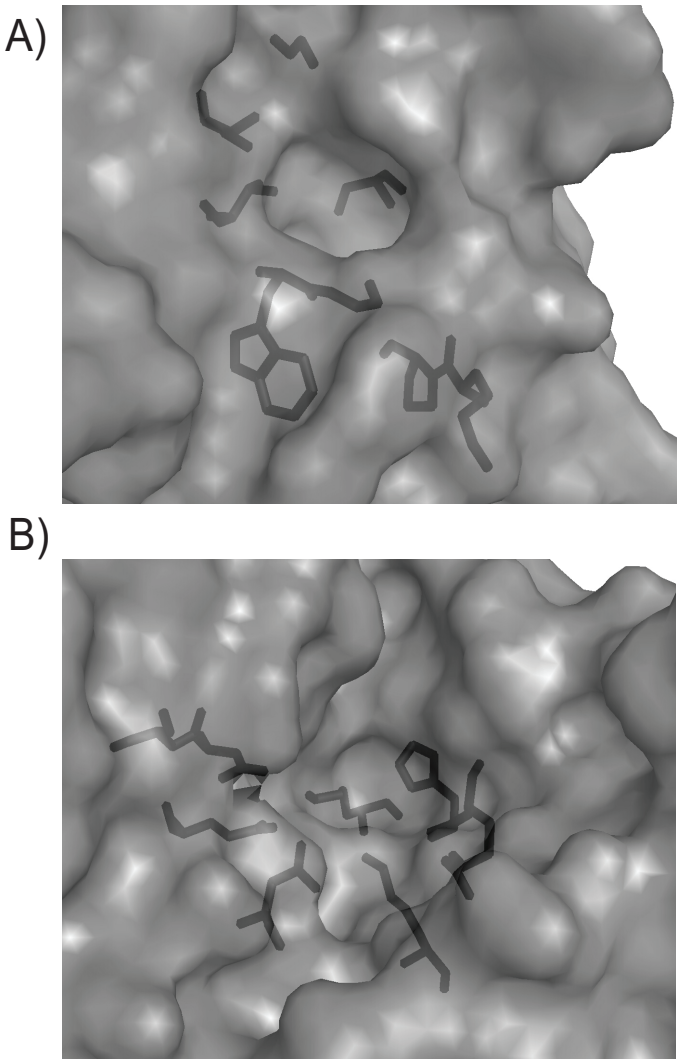

Fig. 4. The 3D structure of the template motif in A) $1a0j$ and B) $1ebh$ shown in the black stick. The protein structure is shown in grey. This figure is generated with Pymol [40].

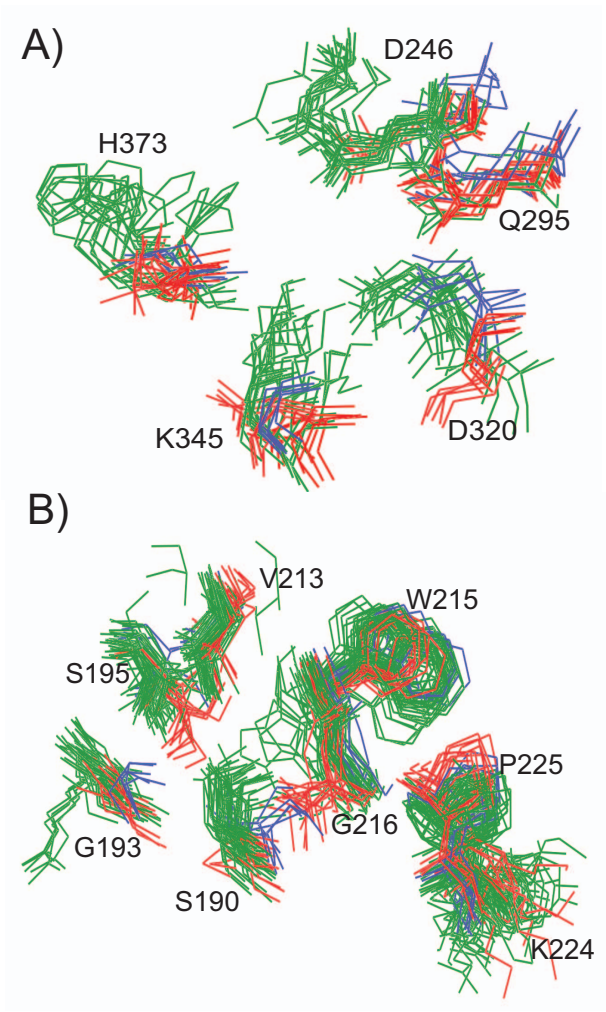

Fig. 5. Superposition of sampled propagated motifs of all proteins in A) serine proteases and B) the enolases where each protein is shown in 5 sampled propagated motifs. The color of each aligned substructure indicates the ligand binding specificity of the protein. It can be seen that propagated motifs of proteins with identical binding specificity are nearby to each other. This figure is generated with Pymol [40].

Overall, UPGMA clusterings on serine proteases and enolases reveal that our ensemble clustering method improves the prediction of protein-ligand binding preferences. It could be a robust tool for flexible protein structure comparisons despite conformational flexibilities of different proteins.

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

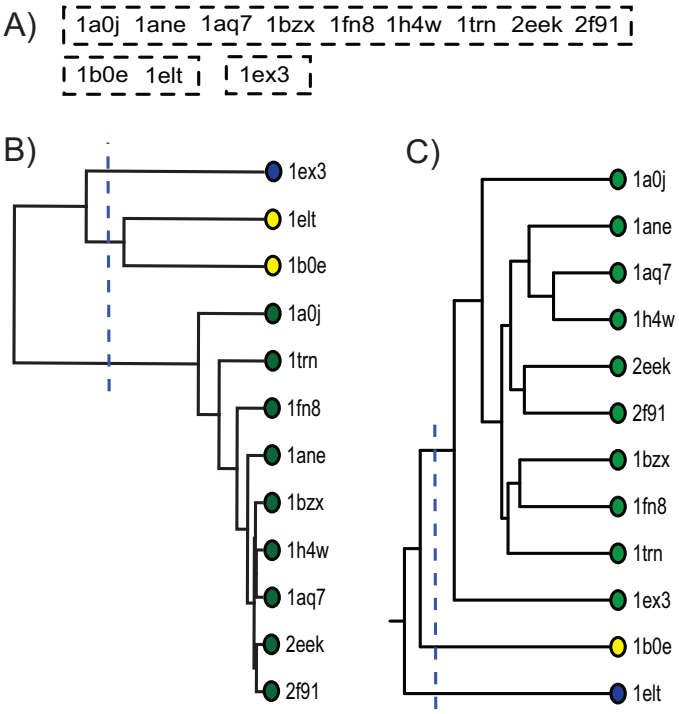

Fig. 6. Comparison of UPGMA clustering of the ensemble method and of FAVA from serine protease structures. A) The ground-truth (EC number) clustering of each subfamily (dotted box) indicates the ligand binding preference of the protein. B) Clustering of the ensemble method using propagated motifs. C) Clustering of frequent regions using FAVA. In both UPGMA trees, the dotted blue line is used to specify the number of subfamilies to generate the coloring of each structure as the prediction in the final clustering vector.

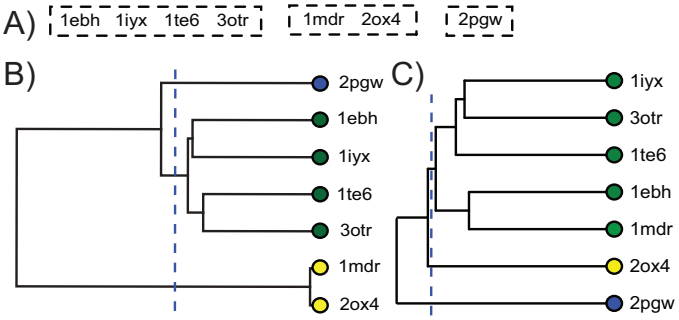

Fig. 7. Comparison of UPGMA clustering of the ensemble method and of FAVA from the enolse superfamily structures. A) The ground-truth (EC number) clustering of each subfamily (dotted box) indicates the ligand binding preference of the protein. B) Clustering of the ensemble method using propagated motifs. C) Clustering of frequent regions using FAVA. In both UPGMA trees, the dotted blue line is used to specify the number of subfamilies to generate the coloring of each structure as the prediction in the final clustering vector.

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.

REFERENCES

[1] R. Nussinov and H. J. Wolfson, "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques," *Proceedings of the National Academy of Sciences*, vol. 88, no. 23, pp. 10 495–10 499, 1991.

[2] C. A. Orengo and W. R. Taylor, "Ssap: sequential structure alignment program for protein structure comparison," *Computer methods for macromolecular sequence analysis*, 1996.

[3] D. Petrey and B. Honig, "Grasp2: visualization, surface properties, and electrostatics of macromolecular structures and sequences." *Methods in enzymology*, vol. 374, pp. 492–509, 2002.

[4] R. B. Russell, "Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution," *Journal of molecular biology*, vol. 279, no. 5, pp. 1211–1227, 1998.

[5] D. H. Bryant, M. Moll, P. W. Finn, and L. E. Kavraki, "Combinatorial clustering of residue position subsets predicts inhibitor affinity across the human kinome," *PLoS computational biology*, vol. 9, no. 6, p. e1003087, 2013.

[6] B. Y. Chen, V. Y. Fofanov, D. H. Bryant, B. D. Dodson, D. M. Kristensen, A. M. Lisewski, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "The mash pipeline for protein function prediction and an algorithm for the geometric refinement of 3d motifs," *Journal of Computational Biology*, vol. 14, no. 6, pp. 791–816, 2007.

[7] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (ce) of the optimal path." *Protein engineering*, vol. 11, no. 9, pp. 739–747, 1998.

[8] L. Holm and C. Sander, "Mapping the protein universe," *Science*, vol. 273, no. 5275, pp. 595–602, 1996.

[9] J.-F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison," *Current opinion in structural biology*, vol. 6, no. 3, pp. 377–385, 1996.

[10] A. R. Poirrette, P. J. Artymiuk, D. W. Rice, and P. Willett, "Comparison of protein surfaces using a genetic algorithm," *Journal of Computer-Aided Molecular Design*, vol. 11, no. 6, pp. 557–569, 1997.

[11] L. Xie and P. E. Bourne, "A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites," *BMC bioinformatics*, vol. 8, no. Suppl 4, p. S9, 2007.

[12] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov, "Molecular shape comparisons in searches for active sites and functional similarity." *Protein Engineering*, vol. 11, no. 4, pp. 263–277, 1998.

[13] K. Kinoshita and H. Nakamura, "Identification of the ligand binding sites on the molecular surface of proteins," *Protein Science*, vol. 14, no. 3, pp. 711–718, 2005.

[14] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang, "Castp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues," *Nucleic acids research*, vol. 34, no. suppl 2, pp. W116–W118, 2006.

[15] T. A. Binkowski and A. Joachimiak, "Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites," *BMC structural biology*, vol. 8, no. 1, p. 45, 2008.

[16] J. Dundas, L. Adamian, and J. Liang, "Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and nad binding proteins," *Journal of molecular biology*, vol. 406, no. 5, pp. 713–729, 2011.

[17] B. Y. Chen and B. Honig, "VASP: A volumetric analysis of surface properties yields insights into protein-ligand binding specificity," *PLoS computational biology*, vol. 6, no. 8, p. e1000881, 2010.

[18] 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 binding," *Journal of molecular biology*, vol. 365, no. 1, pp. 257–273, 2007.

[19] M. Shatsky, R. Nussinov, and H. J. Wolfson, "Flexprot: alignment of flexible protein structures without a predefinition of hinge regions," *Journal of Computational Biology*, vol. 11, no. 1, pp. 83–106, 2004.

[20] J. Konc and D. Janežič, "Probis algorithm for detection of structurally similar protein binding sites by local structural alignment," *Bioinformatics*, vol. 26, no. 9, pp. 1160–1168, 2010.

[21] Y. Ye and A. Godzik, "Multiple flexible structure alignment using partial order graphs," *Bioinformatics*, vol. 21, no. 10, pp. 2362–2369, 2005.

[22] R. Mosca and T. R. Schneider, "Rapido: a web server for the alignment of protein structures in the presence of conformational changes," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W42–W46, 2008.

[23] F. Birzele, J. E. Gewehr, G. Csaba, and R. Zimmer, "Vorolign—fast structural alignment using voronoi contacts," *Bioinformatics*, vol. 23, no. 2, pp. e205–e211, 2007.

[24] M. Menke, B. Berger, and L. Cowen, "Matt: local flexibility aids protein multiple structure alignment," *PLoS computational biology*, vol. 4, no. 1, p. e10, 2008.

[25] J. Vesterstrøm and W. R. Taylor, "Flexible secondary structure based protein structure comparison applied to the detection of circular permutation," *Journal of Computational Biology*, vol. 13, no. 1, pp. 43–63, 2006.

[26] Z. Guo and B. Y. Chen, "Variational bayesian clustering on protein cavity conformations for detecting influential amino acids," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 703–710.

[27] Z. Guo, T. Kuhlengel, S. Stinson, S. Blumenthal, B. Y. Chen, and S. Bandyopadhyay, "A flexible volumetric comparison of protein cavities can reveal patterns in ligand binding specificity," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 445–454.

[28] E. C. Webb *et al.*, *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Academic Press, 1992, no. Ed. 6.

[29] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC Press, 2012.

[30] P. H. Sneath and R. R. Sokal, "Numerical taxonomy," *Nature*, vol. 193, no. 4818, pp. 855–860, 1962.

[31] A. Strehl and J. Ghosh, "Cluster ensembles-a knowledge reuse framework for combining partitionings," in *AAAI/IAAI*, 2002, pp. 93–99.

[32] K. Morihara and H. Tsuzuki, "Comparison of the specificities of various serine proteinases from microorganisms," *Archives of biochemistry and biophysics*, vol. 129, no. 2, pp. 620–634, 1969.

[33] L. Graf, A. Jancso, L. Szilágyi, G. Hegyi, K. Pintér, G. Náray-Szabó, J. Hepp, K. Medzihradszky, and W. J. Rutter, "Electrostatic complementarity within the substrate-binding pocket of trypsin," *Proceedings of the National Academy of Sciences*, vol. 85, no. 14, pp. 4961–4965, 1988.

[34] G. I. Berglund, A. O. Smalas, H. Outzen, and N. P. Willassen, "Purification and characterization of pancreatic elastase from north atlantic salmon (salmo salar)," *Molecular marine biology and biotechnology*, vol. 7, no. 2, pp. 105–114, 1998.

[35] P. C. Babbitt, M. S. Hasson, J. E. Wedekind, D. R. Palmer, W. C. Barrett, G. H. Reed, I. Rayment, D. Ringe, G. L. Kenyon, and J. A. Gerlt, "The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the α-protons of carboxylic acids," *Biochemistry*, vol. 35, no. 51, pp. 16 489–16 501, 1996.

[36] K. Kühnel and B. F. Luisi, "Crystal structure of the escherichia coli rna degradosome component enolase," *Journal of molecular biology*, vol. 313, no. 3, pp. 583–592, 2001.

[37] S. L. Schafer, W. C. Barrett, A. T. Kallarakal, B. Mitra, J. W. Kozarich, J. A. Gerlt, J. G. Clifton, G. A. Petsko, and G. L. Kenyon, "Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the d270n mutant," *Biochemistry*, vol. 35, no. 18, pp. 5662–5669, 1996.

[38] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[39] A.-S. Yang and B. Honig, "An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance." *J Mol Biol*, vol. 301, no. 3, pp. 665–78, Aug. 2000.

[40] W. L. DeLano, "The pymol molecular graphics system," 2002. [Online]. Available: https://www.pymol.org/

[41] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of chemical theory and computation*, vol. 4, no. 3, pp. 435–447, 2008.

[42] H. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans, "Intermolecular forces," *Pullman, B., Ed.; Reidel Publishing Company: Dordrecht*, pp. 331–342, 1981.

[43] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *Journal of Applied physics*, vol. 52, p. 7182, 1981.

[44] S. Nose and M. Klein, "Constant pressure molecular dynamics for molecular systems," *Molecular Physics*, vol. 50, no. 5, pp. 1055–1076, 1983.

[45] B. Hess, "P-lincs: A parallel linear constraint solver for molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 1, pp. 116–122, 2008.