Variational Bayesian Clustering on Protein Cavity Conformations for Detecting Influential Amino Acids

Ziyi Guo, Brian Y. Chen[*]
Department of Computer Science and
Engineering,Lehigh University
Bethlehem,PA,18015
zig312@lehigh.edu
chen@cse.lehigh.com

ABSTRACT

Proteins are large flexible biological molecules and conformational flexibility is a shared challenge in comparisons of protein structure. Many tools have been developed to identify remote homologs in cases where backbone flexibilities are considered. However, these methods require comparisons of structures of more than one proteins, and this is not always available. To assist this process, this paper presents an unsupervised method to predict amino acids that exhibit substantial flexibility to change the binding site when only one protein structure is available. Our method is applied on conformational samples of sequentially nonredundant structures of the serine protease proteins. We observed that influential amino acids can be predicted with high specificities in our whole data set. The results suggest our method as a tool to detect significant side chain motions that affect binding specificity of one protein in the presence of great flexibility.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Algorithms, Design

1. INTRODUCTION

Discovering the elements of enzyme structures that govern the selective recognition of substrates is a shared challenge in many current studies. In computational studies, we have demonstrated that comparisons of proteins with different binding preferences can reveal differences in binding cavities that influence binding specificity. This approach can be unsupervised when paired with statistical models [11, 10], and more effective when combined with homology modeling, to reduce errors from conformational flexibility [19, 20]. Nonetheless, all these methods require comparisons of the structures of two related proteins with different binding preferences, and in many emerging topics of study, such data are not always available.

This paper examines a novel technique for predicting amino acids that influence specificity when only a single structure is available. Our approach is to identify amino acids with side chains whose motion can dramatically alter the shape of the binding cavity. Since these amino acids are in the binding site, and thus unlikely to exhibit evolutionary variability, we hypothesize that their role may be to sterically hinder certain ligands from binding.

Our method analyzes the motion of sets of amino acids within the binding site called motifs, using Boolean operations from Constructive Solid Geometry (CSG) [13] (Figure 1a). Next, using Variational Bayesian Gaussian Mixture Model (VBGMM)[8], a robust algorithm for data clustering, we are able to find clusters of conformations (CCs) of the same protein based on binding site similarity. The CCs represent structural flexibility in the binding cavity and the clusters indicate binding sites in similar conformations. An analysis of these clusters is able to detect influential amino acids that differentiate CCs of the same protein.

We tested our method on sequentially nonredundant protein structures of the serine protease superfamily. On simulations of these proteins, we identified clusters of binding site conformations despite considerable flexibility. Our method analyzed these clusters and detected 5 amino acids that created structural variations of the binding cavity and influenced binding preference of the serine proteases.

2. RELATED WORK

Conformational flexibility is a shared challenge in protein structure comparison. Many comparisons are possible by using rigid transformations to bring atoms from different structures into superposition without considering alternative conformations. This simplicity enables protein structure comparisons via backbone positions [34, 36, 38, 41, 9, 12, 45], distance matrices [25], graphs [18, 39, 48] and geometric surfaces [40, 28, 17, 5, 6] to find similar functional sites. However, without the assumption of rigidity, protein structure analysis of function and binding preference would face a much more general problem because all conformations must then be considered.

In recent years, structure comparison algorithms have used hinges [21, 44], graph structures [29, 50], fragments [32] and dynamic programming [7, 30, 47] to represent protein structures. These techniques use rigid structures with flexible linkers. Most methods are designed to find remote homologs with similar folds that might be overlooked due to

[*]Corresponding author.

BCB'14, September 20–23, 2014, Newport Beach, CA, USA.
Copyright 2014 ACM 978-1-4503-2894-4/14/09 ...$15.00.
http://dx.doi.org/10.1145/2649387.2660837 .

different conformations. However, rigid substructures and flexible linkers do not represent the small motions of side chains inside binding cavities. Thus, conformational flexibility still prevents precise comparisons of functional sites because small motions are not represented by existing flexible comparison methods. In such cases, backbone structures may even be highly similar but side chains motions may generate variations in the binding sites that are overlooked.

3. METHODS

3.1 Overview

Taking conformational samples of one protein structure as input, our method is designed to find CCs in the ligand binding cavity. First, we explain how to define the structural motif: positions of adjacent residues to the ligand surface, and the motif is taken as a feature vector representation of the ligand binding cavity. A set of motifs describe the same amino acids from different conformational samples of binding cavities of the same protein structure, and most geometric features in the motif are highly similar. So these features will increase the dimensionality of the geometric feature space but do not give enough discriminative information in structure to cluster analysis, leading to the motivation for dimension reduction step. Also, dimension reduction makes feature space analysis easier by decoupling the dimensionality of the feature space from the size of the input.

We then describe how we obtain CCs using VBGMM in the reduced feature space. VBGMM is a full Bayesian algorithm for data clustering and has several advantages. First, the Bayesian method solves the singularity problem in maximum likelihood when one Gaussian cluster collapses to only one data point [8]. Second, VBGMM has an automatic sparsity property where clusters with very few members become more and more empty, whereas popular clusters get more and more members. By removing empty clusters, it is able to automatically determine the optimal number of Gaussian components without seeking other techniques with arbitrariness, such as cross validation and information criterion [1].

Given two CCs, $\{CC_i, CC_j\}$, of sampled binding cavities of the same protein, a cavity in CC_i is frequently dissimilar from CC_j, and a set of amino acids is responsible for making their binding structures different. Finally, we discuss how to detect such amino acids automatically.

3.2 Structural Motifs Definition

Formally, we refer to the conformational samples of a protein structure A as $\{A_1, A_2, ..., A_N\}$, and a ligand bound to A as l. In each conformation sample A_i, we keep the coordinates of every atom so that sufficient samples will be provided to represent the structural flexibility of A.

Every sample A_i is aligned onto A using Ska [49], an algorithm for whole-structure alignment. Then, for every atom in l, we generate a sphere with radius 5.0 Å . We use VASP [13], a volumetric analysis tool for rigid region comparisons of protein structures, to compute the CSG union (Figure 1a) of these spheres, S_l, which defines the vicinity of the ligand binding cavity. The amino acid r is added into the motif if these exists at least one conformation where r intersects with S_l. Then, the motif of sample A_i is defined as $m_i = \{AA_1, AA_2, ..., AA_R\}$ where AA_r indicates positions of all atoms in amino acid r. The motif is considered to be close to the ligand binding cavity and its flexibility may influence the shape and structure of the binding site.

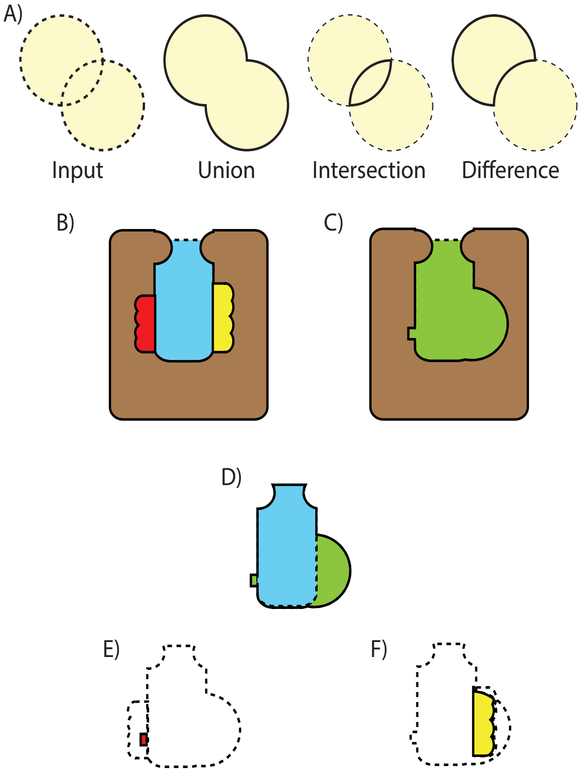

Figure 1: A) CSG operations. Input regions are shown in light yellow with dotted boundaries. Outputs are shown in solid boundaries. B) Input protein A_m (gray) in CC_i with ligand binding cavity a_m (light blue), one amino acid r_1 (red) and another amino acid r_2 (yellow). C) Another input protein A_n (gray) in CC_j with ligand binding cavity a_n (green). D) Superimposition of a_m and a_n based on the whole structure alignment of A_m and A_n. E) A small intersection region α_1 (red) between r_1 and a_n. F) A large intersection region α_2 (yellow) between r_2 and a_n and this indicates that r_2 could be a potentially influential amino acid to make binding cavities in CC_i different from cavities in CC_j.

Given a protein conformational sample A_i, the defined motif m_i is encoded as a geometric feature vector and each feature indicates one direction of the coordinate position of one atom: either x or y or z. The set of feature vectors, $M = \{m_1, m_2, ..., m_N\}$, is a vector representation for motif conformations in binding cavities of A. Each motif m_i defines a point in the high-dimensional space and motifs with similar structures should be nearby where distance is measured by Euclidean distance.

3.3 Dimension Reduction

In this paper, Principal Component Analysis (PCA) is chosen because it is suitable to express small linear motion of backbones or side chains [16]. PCA [27] is a linear dimension reduction procedure to orthogonally convert M, a set of possibly correlated variables, to X, a set of uncorrelated variables named principal components (PCs) with the number of PCs usually less than the dimension of original variables. Here, we just retain the top 2 PCs because they are more significant than the others and capture a large percent of original variances (See section 5.3). Thus, PCA is capable of largely reduce the dimension by mapping motifs into a $2D$ space.

3.4 Variational Bayesian Clustering

Gaussian Mixture Model (GMM) is a probabilistic model written as a linear combination of Gaussians in the form $p(x) = \sum_k \pi_k N(x_n|\mu_k, \Lambda_k^{-1})$ where π_k is the mixture coefficient, μ_k is the mean and Λ_k is the precision matrix. In the Bayesian approach, we consider conjugate priors for model parameters: Dirichlet distribution for matrix coefficients, Gaussian distribution for mean and Wishart distribution for precision matrix. Using mean field theory [35] for Bayesian inference, these parameters can be accurately estimated with an iterative method that is similar to Expectation Maximization (EM) in the maximum likelihood framework. In this section, the input is $X = \{x_1, x_2, ..., x_N\}$, the vectors in the reduced feature space. After convergence, the mixture coefficients indicate the cluster membership of each data point, forming CCs of sampled binding cavities. More details of VBGMM can be found in [8].

3.5 Volumetric Similarity Computation

Given the conformational samples $\{A_1, A_2, ..., A_N\}$ of protein structure A and the binding ligand l, we define the shape of the sampled ligand binding cavities as $\{a_1, a_2, ..., a_N\}$ following our earlier work [22], leading to the volumetric representation for surface properties of binding cavities [13]. Then, we measure the volumetric distance, $D(a_i, a_j)$, between two protein sampled cavities using Tanimoto Coefficient [26]:

$$D(a_i, a_j) = 1 - \frac{|a_i \cap a_j|}{|a_i \cup a_j|} \tag{1}$$

where $|x|$ indicates the volume of a solid region x. The volume is calculated by the Surveyors Formula [42] and has been used in some of our earlier works [13, 11, 10]. Here, the less $D(a_i, a_j)$ is, the more volumetrically similar these two binding cavities are.

Note that, after using VBGMM clustering on vector representation of protein motifs, $\{a_i, a_j\}$ could come from the same CC with similar motif structures or different CCs with dissimilar structures. Here, we expect that sampled binding cavities in the same CC will also be nearby in volumetric distance.

3.6 Influential Amino Acid Detection

Given two clusters CC_i and CC_j, we consider one conformation A_m in CC_i and another conformation A_n in CC_j of protein structure A. For one amino acid r in the motif of A_m, we define the molecular surface of r as $m(r)$ using the rolling probe algorithm [14] with the standard radius size of 1.4 Å . We say that r makes the cavity of A_m, called a_m, different from the cavity of A_n, called a_n if $m(r)$ has a nonempty intersection region α with a_n (Figure 1 f). In this case, α is not solvent accessible to a_m but accessible to a_n.

Amino acids that radically change the structures of cavities alter binding shape. To evaluate how influential of a given amino acid r, we compute $INT_r(CC_i, CC_j)$, the median intersection volume between r of A_m in CC_i and binding cavity of A_n in CC_j, for all pairs of m and n. Nontrivial value of $INT_r(CC_i, CC_j)$ indicates that r frequently makes cavity structures of CC_i different from CC_j and r is defined as *influential amino acid*.

4. DATA SET CONSTRUCTION

Serine Protease Superfamily:
Chymotrypsins: 1ex3
Elastases: 1b0e, 1elt
Trypsins: 1a0j, 1ane, 1aq7, 1bzx, 1fn8, 1h4w, 1trn, 2eek, 2f91

Figure 2: PDB codes of structures used.

4.1 Protein Selection

To demonstrate that our method is capable of effectively finding CCs in binding cavity conformations, protein structures in three subfamilies, the trypsin, chymotrypsin and elastase, of the serine protease superfamily were selected.

The serine protease structures in our dataset were downloaded from the Protein Data Bank [4] on 6.21.2011. These 676 structures, which were selected based on their enzyme classification (EC), were then filtered to remove mutants and structures with disordered regions were removed. Next, the structures were filtered to maintain less than 90% pairwise sequence identity, with a preference to keep structures associated with publications. Technical problems with simulation prevented proteins 8gch and 1aks from being added. From the remaining 12 structures, ions, waters, and non-protein atoms were removed. Since hydrogens were available in only some structures, all hydrogens were removed for uniformity.

All structures in our data set were aligned to bovine gamma-chymotrypsin (pdb:8*gch*) because of its availability of the bound ligand which was used to generate sampled binding cavities through rigid structure representations.

4.2 Protein Structure Simulation

Conformational samples for each structure in the data set were generated using GROMACS 4.5.4 [24]. Structures were prepared for simulation by centering them inside a cubic waterbox with fully periodic boundary conditions that was populated using the 3-point solvent model SPC/E [2]. The waterbox was sized to contain the protein with at least 10 Å between the protein and the nearest part of the box. Charge balanced sodium and potassium ions were then added to the solvent at a low concentration ($< 0.1\%$ salinity). After

Table 1: Protein Motif Examples

PDB	Motifs
1b0e	H57 I138 L160 G185 S189 G190 C191 Q192 G193 D194 S195 G196 T213 S214 F215 V216 C220 K224 T226
1h4w	H57 I138 L158 D189 S190 C191 Q192 S195 V213 S214 W215 G216 G219 C220 R224 P225 Y228
1ex3	G142 L160 W172 V188 S189 S190 C191 G193 D194 S195 V213 S214 W215 G216 S217 S221 T224 P225 Y228

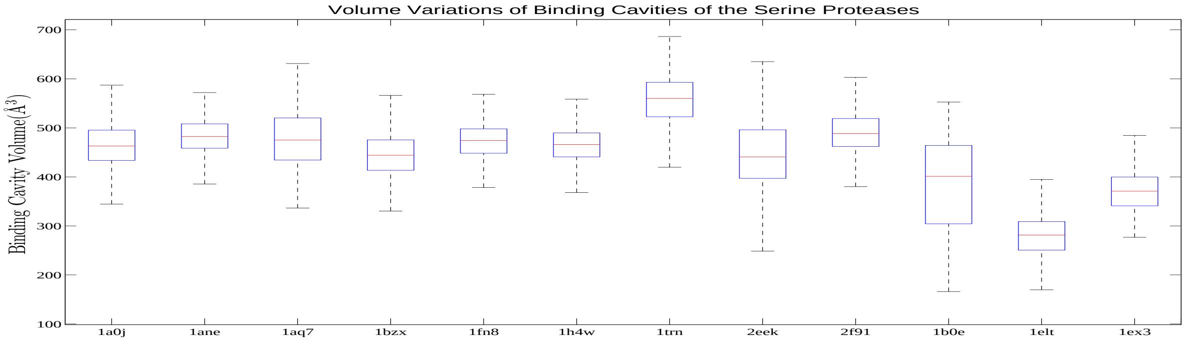

Figure 3: Aggregate variations in sampled cavity volume in our whole data set. All protein cavity samples varied considerably.

preparation, energy minimization was performed on the entire system using a steepest descent algorithm. Isothermal-Isobaric (NPT) equilibration was performed in four 250 picosecond steps to permit temperature and pressure equilibration prior to the primary simulation. Over the 1000 picosecond minimization period, at 1000 $kJ/(mol * nm)$, each equilibration step reduced the position restraint force by 250 $kJ/(mol * nm)$. For the primary NPT simulation, backbone position restraints were released. System energies were generated at the start of the equilibration phase. Initially, temperature was set to 300 Kelvin and pressure was set to 1 bar. Temperature coupling was calculated using the Nosé-Hoover thermostat [2], and the Parrinello-Rahman algorithm [37, 33] was used for pressure coupling. P-LINCS [23] was used to update bonds, and electrostatic interaction energies were calculated by particle mesh Ewald summation (PME) [24]. P-LINCS and PME were chosen for their parallel efficiency. 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. OpenMPI was used for inter-node and inter-process communication, on multiple 16 core nodes. PME distribution was automatically selected by GROMACS. After the simulation was completed, 600 conformational samples were selected at uniform intervals for our data set.

5. RESULTS

First, we demonstrate how substantially volumes of ligand binding cavities vary over the simulation, leading to the motivation to find CCs through unsupervised methods. Then, we demonstrate the clustering results on sampled proteins using VBGMM, and showed volumetric similarity within the same CC. Finally, we demonstrate the prediction of influential amino acids with high specificity.

5.1 Ligand Binding Cavities Vary Considerably

In Figure 3, we observe significant volume variations of binding cavities in conformational samples of protein structures in our data set. Specifically, we observe that the trypsin cavity volumes ranged from 248 $\mathring{A}^3$ to 692 $\mathring{A}^3$, the chymotrypsin cavity volumes ranged from 276 $\mathring{A}^3$ to 568 $\mathring{A}^3$ and the elastase cavity volume ranged from 126 $\mathring{A}^3$ to 552 $\mathring{A}^3$, despite the general idea that chymotrypsin cavities are larger so that to accommodate aromatic side chains [31] while elastase cavities are smaller to accommodate small amino acids, such as valine and alanine [3]. The volume of sampled binding cavities varied because of side chain motions and smaller backbone motions, which enlarged, shrank or separated the structures of the cavity.

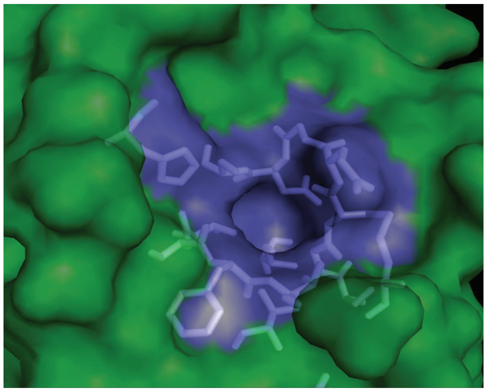

Figure 4: $3D$ structure of $S1$ pocket of porcine pancreatic elastse (pdb:1b0e) and the binding motif is shown in purple. This figure is generated with Pymol [15].

From these observations, we found that the flexibility of the serine proteases creates considerable variations among sampled binding cavities of the same protein, preventing accurate prediction of protein binding preference. Cluster analysis, which is able to find similar binding cavity conformations, may be a solution to avoid this problem.

5.2 Generating Binding Cavity Motifs

In Table 1, we show the motifs of one representative protein structure in each subfamilies. We observe that these motifs are highly similar, and we believe that the common amino acids jointly construct residues that are potentially influential to $S1$ pocket of the serine protease structures. The Figure 4 shows $3D$ structure of the motif in one conformational sample of porcine pancreatic elastase(pdb: 1b0e).

5.3 Clustering Binding Cavities

Figure 5 illustrate the average percentage of total variances of top 10 PCs over all the 12 serine protease conformations. It is obvious that top 2 PCs are much more significant than other PCs and account for a large percent of the total variances.

Figure 6 illustrate a VBGMM demo on the protein of porcine pancreatic elastse (pdb:1b0e). It is clear that after model convergence, there are only four non-empty Gaussian components while the rest are automatically dropped since VBGMM is able to trade-off between fitting the data and the complexity of the model, resulting in automatic detection of the optimal number of clusters.

Figure 7 shows volumetric similarities between all sampled protein cavities of porcine pancreatic elastase (pdb:1b0e). We observe that protein cavities in the same cluster are highly similar in the volumes. Besides, in Figure 7, we find infrequently-occurred extra-cluster similarities (e.g. between green cluster and teal cluster). These similarities are generated by arbitrariness of data clustering, and, following these extra-cluster volumetric similarities, we usually find geometric similarities between corresponding CCs. For example, we find that the green cluster and the teal cluster are highly close in Figure 6. Similarly, we can always found intra-cluster volumetric similarities for other protein structures in our data set.

5.4 Detecting Influential Amino Acids

Changes in the backbone and side chain positions create structural flexibility of ligand binding cavities, leading to different CCs in conformational samples. To evaluate how different amino acids create such changes, for protein A, we computed $INT_r(CC_i, CC_j)$ for all residues r in all sampled structures A_m in one cluster CC_i and all binding cavities in another cluster CC_j.

Most amino acids exhibited very small median intersection volume, indicating little influence on structural flexibility between two different CCs. However, some amino acids were identified with frequently large intersections. Figure 8 shows the median intersection volume between amino acids from conformational samples in the red cluster and binding cavities from conformational samples in the teal cluster of porcine pancreatic elastase (pdb:1b0e). Only serine 195 exhibited a large median intersection of 23.5 $\mathring{A}^3$. One example, illustrated in Figure 9 depicts that influential serine 195 from one conformational sample in CC_{RED} intersected a large volume of 44.6 $\mathring{A}^3$ with cavity from one conformational sample in CC_{TEAL}. However, another amino acid of the same conformational sample, lysine 224, intersected a very small volume of 1.4 $\mathring{A}^3$ despite its adjacency to the binding ligand.

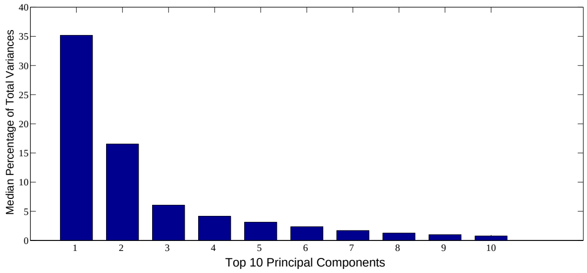

Figure 5: Average percentage of total variances of top 10 PCs over all 12 structures in our whole set.

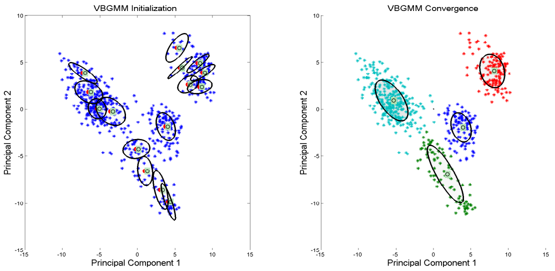

Figure 6: VBGMM demo applied on the protein of porcine pancreatic elastse (pdb:1b0e). Left Figure shows the initialization plot using K-means algorithm in the PCA mapped feature space where the prior number of clusters is 15. The ellipses denote the one standard deviation density contours for each of the clusters and each red point marked as ∘ corresponds to the center of each cluster. Right Figure shows the data plot after convergence where those empty clusters are not plotted and the colors denote cluster labels.

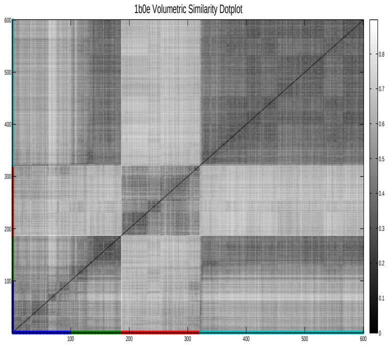

Figure 7: Volumetric similarity dotplot of porcine pancreatic elastase (pdb:1b0e). The axis (x or y) shows the sequences of sampled binding cavities of 1b0e and colors along the axes denote the cluster labels shown in Figure 6.

As illustrated on Table 2, influential amino acids are predicted on all serine proteases. We validated these amino acids against the experimental literature, and observed that most of these predictions play notable roles in specificity and function. For example, in atlantic salmon trypsin (pdb: 1a0j), serine 195 is the nucleophilic serine [43]. In salmon elastase (pdb: 1elt), valine 216 creates steric hindrance in the S1 binding site that prevents larger molecules from binding [46]. In total, six predictions were made from 220 amino acids. 5 predictions were correct when validated against experimental findings, and one, tryptophan 215 was not functional to our knowledge. It appears that because W215 is adjacent to V216, it also overlaps with cavity conformations in many cases.

6. CONCLUSION

We have presented a technique for predicting amino acids that exhibit substantial flexibility within the ligand binding sites of an individual protein structure. Unlike existing methods, this approach does not require comparisons

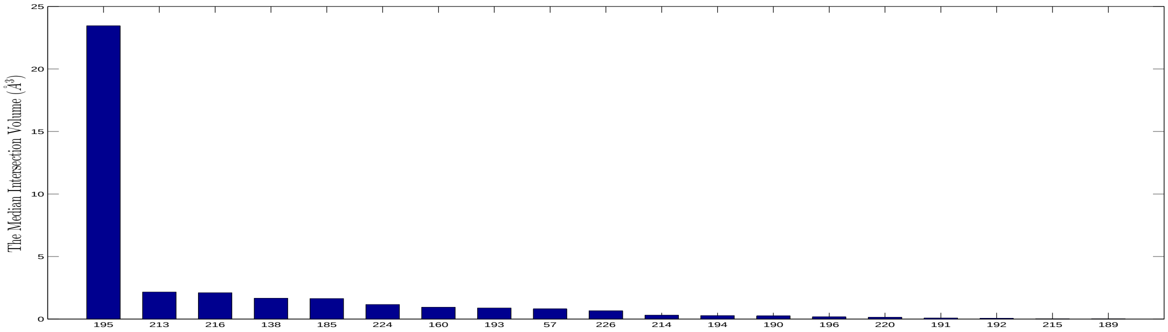

Figure 8: Aggregate intersection volumes of amino acids from conformational samples in CC_{RED} with cavities from conformational samples in CC_{TEAL} of porcine pancreatic elastase (pdb:1b0e). The subscript color of CCs can be found in Figure 6.

Table 2: Amino acid prediction on all paris of CCs of all serine protease proteins in our data set.

PDB code	Predicted Amino Acids
1a0j	**S195**
1ane	None
1aq7	None
1bzx	None
1fn8	None
1h4w	None
1trn	W215, **V216**
2eek	None
2f91	None
1b0e	**S195**
1elt	**V216**, **S195**
1ex3	None

Amino acids are selected when the median intersection volume is greater than 20 Å^3 and the bolded amino acids are validated in the literatures.

against a similar structure with different binding preferences. Instead, it relies on the inference that amino acids in the binding site that are large enough to alter the apparent shape of the site must be evolutionarily selected for their purpose and that they will sterically hinder some ligands at the binding site.

Our findings support this claim on the serine proteases, where we identified 5 amino acids that influence specificity and function with 1 false positive prediction. While a larger number of influential amino acids can be identified with comparative methods, they rely on the presence of at least two protein structures, and the method described here does not.

Applications of our method exist in contexts where the identification of influential amino acids is required and only a single protein structure exists. In such cases, experimental procedures are time consuming and expensive, and it is thus critical to avoid false positives. The predictions identified here, on a limited set of proteins, indicate that it may be possible to find such amino acids without many extraneous predictions, and thus that a flexibility study of individual structures may still reveal elements of structure that influence binding.

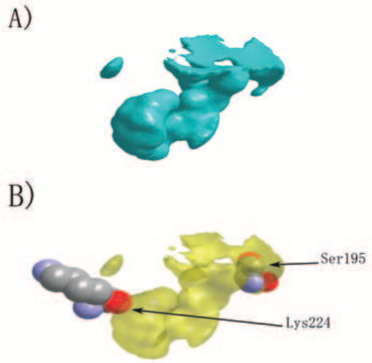

Figure 9: A) The ligand cavity from one conformational sample in CC_{TEAL} of porcine pancreatic elastase (pdb:1b0e). B) The positions of serine 195 and lysine 224 from one conformational sample in CC_{RED} of porcine pancreatic elastase (pdb:1b0e).

Acknowledgment

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

7. REFERENCES

[1] Akaike, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*, 6 (1974), 716–723.

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

[3] Berglund, G. I., Smalas, A. O., Outzen, H., and Willassen, N. P. Purification and characterization of pancreatic elastase from north atlantic salmon (salmo salar). *Molecular marine biology and biotechnology 7*, 2 (1998), 105–114.

[4] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research 28*, 1 (2000), 235–242.

[5] Binkowski, T. A., Adamian, L., and Liang, J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *Journal of molecular biology 332*, 2 (2003), 505–526.

[6] Binkowski, T. A., and Joachimiak, A. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC structural biology 8*, 1 (2008), 45.

[7] Birzele, F., Gewehr, J. E., Csaba, G., and Zimmer, R. Vorolign—fast structural alignment using voronoi contacts. *Bioinformatics 23*, 2 (2007), e205–e211.

[8] Bishop, C. M., et al. *Pattern Recognition and Machine Learning*, vol. 1. springer New York, 2006.

[9] Bryant, D. H., Moll, M., Finn, P. W., and Kavraki, L. E. Combinatorial clustering of residue position subsets predicts inhibitor affinity across the human kinome. *PLoS computational biology 9*, 6 (2013), e1003087.

[10] Chen, B., and Bandyopadhyay, S. A statistical model of overlapping volume in ligand binding cavities. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on* (2011), IEEE, pp. 424–431.

[11] Chen, B. Y., and Bandyopadhyay, S. VASP-S: A volumetric analysis and statistical model for predicting steric influences on protein-ligand binding specificity. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on* (2011), IEEE, pp. 22–29.

[12] Chen, B. Y., Fofanov, V. Y., Bryant, D. H., Dodson, B. D., Kristensen, D. M., Lisewski, A. M., Kimmel, M., Lichtarge, O., and Kavraki, L. E. The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3d motifs. *Journal of Computational Biology 14*, 6 (2007), 791–816.

[13] Chen, B. Y., and Honig, B. VASP: A volumetric analysis of surface properties yields insights into protein-ligand binding specificity. *PLoS computational biology 6*, 8 (2010), e1000881.

[14] Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science 221*, 4612 (1983), 709–713.

[15] DeLano, W. L. The pymol molecular graphics system.

[16] Dhanik, A., and Kavraki, L. E. Protein–ligand interactions: Computational docking. *eLS* (2001).

[17] Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., and Liang, J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic acids research 34*, suppl 2 (2006), W116–W118.

[18] Gibrat, J.-F., Madej, T., and Bryant, S. H. Surprising similarities in structure comparison. *Current opinion in structural biology 6*, 3 (1996), 377–385.

[19] Godshall, B. G., and Chen, B. Y. Improving accuracy in binding site comparison with homology modeling. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on* (2012), IEEE, pp. 662–669.

[20] Godshall, B. G., Tang, Y., Yang, W., and Chen, B. Y. An aggregate analysis of many predicted structures to reduce errors in protein structure comparison caused by conformational flexibility. *BMC structural biology 13*, Suppl 1 (2013), S10.

[21] Gunasekaran, K., and Nussinov, R. 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 365*, 1 (2007), 257–273.

[22] Guo, Z., Kuhlengel, T., Stinson, S., Blumenthal, S., Bandyopadhyay, S. R., and Chen, Y. B. Flexible volumetric comparison of protein cavities can reveal patterns in ligand binding specificity. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine* (2014), ACM, p. in press.

[23] Hess, B. P-LINCS: a parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation 4*, 1 (Jan. 2008), 116–122.

[24] Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation 4*, 3 (Mar. 2008), 435–447.

[25] Holm, L., and Sander, C. Mapping the protein universe. *Science 273*, 5275 (1996), 595–602.

[26] Jaccard, P. The distribution of the flora in the alpine zone. 1. *New phytologist 11*, 2 (1912), 37–50.

[27] Jolliffe, I. *Principal Component Analysis*. Wiley Online Library, 2005.

[28] Kinoshita, K., and Nakamura, H. Identification of the ligand binding sites on the molecular surface of proteins. *Protein Science 14*, 3 (2005), 711–718.

[29] Konc, J., and Janežič, D. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics 26*, 9 (2010), 1160–1168.

[30] Menke, M., Berger, B., and Cowen, L. Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology 4*, 1 (2008), e10.

[31] Morihara, K., and Tsuzuki, H. Comparison of the specificities of various serine proteinases from microorganisms. *Archives of biochemistry and biophysics 129*, 2 (1969), 620–634.

[32] Mosca, R., and Schneider, T. R. Rapido: a web server for the alignment of protein structures in the presence of conformational changes. *Nucleic acids research 36*, suppl 2 (2008), W42–W46.

[33] Nose, S., and Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics 50*, 5 (1983), 1055–1076.

[34] Nussinov, R., and Wolfson, H. J. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proceedings of the National Academy of Sciences 88*, 23 (1991), 10495–10499.

[35] Opper, M., and Saad, D. *Advanced Mean Field Methods: Theory and Practice*. MIT press, 2001.

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

[37] Parrinello, M., and Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics 52* (1981), 7182.

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

[39] Poirrette, A. R., Artymiuk, P. J., Rice, D. W., and Willett, P. Comparison of protein surfaces using a genetic algorithm. *Journal of Computer-Aided Molecular Design 11*, 6 (1997), 557–569.

[40] Rosen, M., Lin, S. L., Wolfson, H., and Nussinov, R. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Engineering 11*, 4 (1998), 263–277.

[41] Russell, R. B. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *Journal of molecular biology 279*, 5 (1998), 1211–1227.

[42] Schaer, J., and Stone, M. Face traverses and a volume algorithm for polyhedra. In *New Results and New Trends in Computer Science*. Springer, 1991, pp. 290–297.

[43] Schechter, I., and Berger, A. On the size of the active site in proteases. i. papain. *Biochemical and biophysical research communications 27*, 2 (1967), 157–162.

[44] Shatsky, M., Nussinov, R., and Wolfson, H. J. Flexprot: alignment of flexible protein structures without a predefinition of hinge regions. *Journal of Computational Biology 11*, 1 (2004), 83–106.

[45] Shindyalov, I. N., and Bourne, P. E. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein engineering 11*, 9 (1998), 739–747.

[46] Shotton, D., and Watson, H. Three-dimensional structure of tosyl-elastase. *Nature 225* (1970), 811–816.

[47] Vesterstrøm, J., and Taylor, W. R. Flexible secondary structure based protein structure comparison applied to the detection of circular permutation. *Journal of Computational Biology 13*, 1 (2006), 43–63.

[48] Xie, L., and Bourne, P. E. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC bioinformatics 8*, Suppl 4 (2007), S9.

[49] Yang, A.-S., and Honig, B. 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. *Journal of molecular biology 301*, 3 (2000), 665–678.

[50] Ye, Y., and Godzik, A. Multiple flexible structure alignment using partial order graphs. *Bioinformatics 21*, 10 (2005), 2362–2369.