A Statistical Model of Overlapping Volume in Ligand Binding Cavities

Brian Y. Chen[1]
Dept. of Computer Science and Engineering
Lehigh University
Bethlehem, PA, USA
chen@cse.lehigh.edu

Soutir Bandyopadhyay
Dept. of Mathematics
Lehigh University
Bethlehem, PA, USA
sob210@lehigh.edu

Abstract

*Understanding and predicting protein-ligand binding preferences is an essential aspect of research in many fields, especially drug design. To assist in this effort, this paper presents VASP-I (*V*olumetric* A*nalysis of* S*urface* P*roperties for* I*ntersections), a statistical model for estimating the probability that a set of cavities exhibit the same conserved region, and may thus have the same binding preferences. We applied this method to analyze ligand binding cavities of sequentially nonredundant structural representatives of the serine protease and enolase superfamilies. On these datasets VASP-I correctly distinguished sets of cavities with identical binding preferences from other sets with varying binding preferences. These results indicate that it can be possible to predict binding cavities that exhibit different binding preferences, even when the biochemical mechanism is unknown.*

1. Introduction

Identifying the elements of protein structure that influence protein-ligand binding specificity is a crucial goal in many fields of protein engineering and drug design. A guiding principle in these fields is that subtle variations on the same binding cavity can enable families of proteins to catalyze the same reaction on different preferred substrates. When observed in protein structures, similarities and variations of this nature can be used to develop hypotheses connecting elements of protein structure to their effect on binding preferences. For example, the inhibition of ATP binding sites in protein kinases is modulated by a binding site variation caused by a gatekeeper residue [1] that hinders large inhibitors, causing drug resistance [2]. Testing such hypotheses can reveal molecular mechanisms that underlie protein-ligand recognition [3], verify mechanisms for drug resistance [2], and suggest a molecular basis for the organization of larger biological systems, such as the selective adhesion of cells [4].

The observations that drive this type of hypothesis development depend on expert knowledge in structural biology and tools for the three dimensional visualization of protein structures (e.g. [5]–[7]). Such methods are essential for elucidating the often subtle variations (e.g. [8]) that can affect specificity, but they become increasingly impractical to use in a comprehensive manner as crystallographic data and structure models proliferate. But fully investigating the larger space of emerging data should enhance our understanding of the elements of protein structure that influence specificity. Homology modeling in particular (e.g. [9], [10]) creates an opportunity to consider the structural impact of a large space of conformational and mutational changes that create differences in cavity shape and thereby alter ligand binding preferences.

To accelerate the analysis of binding cavities and to quantify the degree of cavity shape variation necessary to alter specificity, we present VASP-I (**V**olumetric **A**nalysis of **S**urface **P**roperties for **I**ntersections). VASP-I is a statistical model trained on the degree of volumetric similarity between aligned binding cavities with identical binding preferences. Once trained, VASP-I estimates the probability that the volumetric similarity between an test set of cavities could be consistent with similarities amongst training set cavities. We hypothesize that when this probability is improbably low, the test set cavities are unlikely to exhibit similar binding preferences.

We tested VASP-I on two sequentially nonredundant families of protein structures: the serine proteases and the enolase superfamilies. In multi-part cross-validation, VASP-I was able to distinguish the degree of volumetric similarity between ligand binding cavities with similar binding preferences, from those with different binding preferences. Such functionalities

1. Corresponding Author

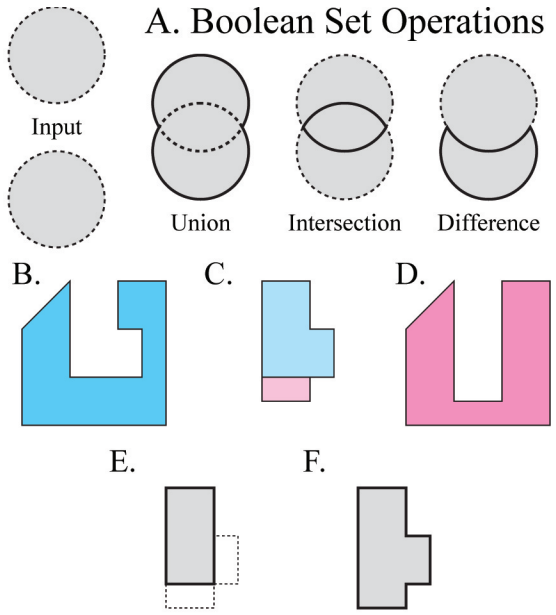

Fig. 1. **A diagram of Boolean set operations (A). Aligned proteins with distinctive cavities (B,D). Overlapping cavities (C). The Boolean intersection (E) and union (F) of the cavities, which is used to compute volumetric similarity.**

point to applications in the large scale classification of binding cavities with identical function, based on differing binding preferences, as well as the detection of binding cavities with potentially novel binding preferences.

2. Related Work

VASP-I builds on a new approach to protein structure comparison based on volumetric similarities and differences in ligand binding cavities [11]. This approach varies considerably from most existing methods, which represent protein structures using points (point-based representations) and surfaces (surface-based representations) in three dimensions. For both point- and surface-based methods, statistical models have been developed for estimating the significance of geometric similarity for differing applications. In contrast, VASP-I presents a new statistical model for volume-based comparison methods, and the first statistical model of overlapping volume.

Point-based representations have notable strengths in comparison efficiency. The least-squares alignment of points in space [12] enables structure comparison software to rapidly consider thousands of atomic superpositions in a search for the alignment of two or more protein structures with greatest geometric similarity [6], [13]–[16]. Other approaches to point-based structure alignment, which employ distance matrices [17] and geometric graphs [18], [19] are also extremely efficient. These alignment methods inspired the design of newer algorithms for the flexible alignment of protein structures [20]–[22], and fuel the ongoing exploration of the space of protein folds [23]. As more protein structures become available, the topology of this space, mapped with structure comparison algorithms, appears to be evolving from earlier fold-based clusterings [24] to a more continuous space of variations [19], [25], [26].

A second class of point-based methods search for functionally related binding sites. Methods of this type encode only the atoms of the binding site itself, sometimes referred to as a *motif* [27]–[29], in order to identify similar functional sites independent of protein fold. One of the major challenges in this subfield has been the design of effective motifs that sensitively align with all functionally related binding sites, while specifically avoiding functionally unrelated sites. To design more effective motifs, supporting algorithms can select atoms that yield more accurate alignments [27], [30], [31], to integrate geometric data from multiple structures [32]–[34], and the integration of empty spaces inside binding sites [35], [36]. As a result of these developments, motif comparison algorithms can be extremely accurate point-based methods for identifying proteins that catalyze the same reaction [33].

Surface-based methods use surfaces or surface patches to represent solvent-accessible shape [37], [38]. The surface itself is often described with triangular meshes [39], [40], three dimensional grids [41], alpha shapes [42]–[44], and spherical harmonics [45]–[47]. Surface representations have been applied for the comparison of protein structures [39], [40] and electrostatic potentials [48], as well as in hybrid representations that combine point-based and surface-based information [35], but they can also be used to predict the location of binding sites [42], [49]–[51] and hot spots [52].

Statistical modeling is a critical aspect of both point- and surface-based methods, because it enables an quantitative separation between signal and noise: Empirical [43], parametric [28], [53], and nonparametric [54] models can identify pairs of functional sites that are too similar to have occurred by random chance. Parametric models can also identify variations in protein ligand binding cavities that are large enough to influence specificity [55]. In contrast to these existing models, VASP-I models volumetric similarity between cavities with identical binding preferences. This model enables

the identification of cavities that are too dissimilar to have the same binding preferences as those on which the model was trained.

3. Methods

By training the VASP-I model on the degree of volumetric similarity typically identified among aligned binding cavities with identical binding preferences, we hypothesize that in groups of cavities with volumetric similarity low enough to be statistically significant there exist cavities with different binding preferences. Here, we first describe how we compute volumetric similarity, how we train the statistical model, and how statistical significance is determined.

3.1. Computing Volumetric Similarity

VASP-I measures volumetric similarity using techniques in computational solid geometry (CSG), developed in earlier work [11]. CSG, developed originally for accurately representing machine parts [56], enables VASP-I to compute Boolean intersection ($\cap$) and union ($\cup$) regions among an input set of cavities (Figure 1A). Using the Surveyor's Formula [57], VASP-I can compute the volume $v(r)$ occupied by any region r. These techniques are detailed in earlier work [11].

Given a set of aligned binding cavities C, we now formally define volumetric similarity $d(C)$ using the Jaccard index:

$$d(C) = \frac{v\left(\bigcap_{i=1}^{k} c_i\right)}{v\left(\bigcup_{i=1}^{k} c_i\right)}$$

Where $c_1, c_2, ...c_k$ are the k members of C. The geometric interpretation of a set of aligned cavities with high volumetric similarity is that they overlap closely, and thus have very similar shape. In contrast, cavities with low volumetric similarity overlap poorly.

3.2. Statistical Model of Volumetric Similarity

VASP-I employs a hypothesis testing framework. Underlying this framework is the assumption that aligned cavities with identical binding preferences will exhibit a *large* degree of volumetric similarity. Conversely, we assume that aligned cavities with differing binding preferences exhibit an *unusually small* degree of volumetric similarity, relative to cavities with identical binding preferences. Beginning with these assumptions, and an input set of k aligned cavities C, our null hypothesis is that $d(C)$ is *large*. The alternative hypothesis is that $d(C)$ is *unusually small*. Because the null hypothesis and the alternative hypothesis are logical complements, only one of these assumptions can hold.

VASP-I tests the null hypothesis by first assuming that it holds for C, and then estimating the probability p of randomly observing another set of k cavities C' with $d(C') \leq d(C)$. If the probability of observing another set of aligned cavities with less volumetric similarity is improbably low (typically .05) then it is hard to reasonably continue assuming that the null hypothesis holds. Under these circumstances, we reject the null hypothesis in favor of the alternative hypothesis, that $d(C)$ is low because the cavities in C have different binding preferences. We can interpret this decision biologically from our underlying assumptions: If the degree of volumetric similarity between the k input cavities is unusually low relative to the degree of volumetric similarity typically observed between cavities with identical binding preferences, then we take this as evidence that the input cavities are unlikely to have identical binding preferences. Rather than being a statement of fact, the rejection of the null hypothesis represents a prediction based on quantified evidence gathered during the training phase.

To perform this prediction, we must estimate the probability p, which requires us to train the statistical model. Our training set, T, consists of n aligned cavities from proteins known to exhibit identical binding preferences. There must be more than k such cavities, and ideally quite a few more. For every combination t, composed of k cavities selected from T, we compute the volumetric distance $d(t)$. These combinations yield $\binom{n}{k}$ volumetric distances to train the model, which is intended to represent the range of volumetric distances to be expected in any set of k binding cavities with preferences identical to those in T. These data are represented in a frequency distribution D (See Figure 4A).

It happens that the shape of D tightly fits a $log\text{-}normal$ distribution, as demonstrated in Section 4. For this reason, we can use it to estimate the probability p of observing k cavities, C', with $d(C')$ less than that of our input set, $d(C)$, and specificity identical to cavities in T. We can make this estimation by approximating the essential parameters of the $log\text{-}normal$ distribution: μ and σ, which are the mean and standard deviation for the log-transformed distribution respectively. We estimate those parameters by simply taking calculating the sample mean and standard deviation of the log-transformed data. Finally, we estimate p using equation 1. We estimate p to be the proportion

$$p(d(F') \geq d(F)) = 1 - \Phi(\frac{\log d(F) - \mu}{\sigma}) \approx 1 - \Phi(\frac{\log d(F) - \overline{x}}{s}). \quad (1)$$

Fig. 2. **Computing the p-value using the best fitting log-normal distribution.**

of the volume under the log-normal curve to the left of $d(C)$, relative to the total volume under the curve ($x \geq 0$).

The advantage of fitting the $log\text{-}normal$ distribution to D is that p can be smoothly estimated without discretizing effects from samples in the training data (e.g. individual t, described above). Also, if we assume that the fitted $log\text{-}normal$ distribution accurately estimates the underlying probability p, then we can use the $log\text{-}normal$ distribution to extrapolate p values beyond that of the smallest $d(t)$ observed on our training set. This kind of extrapolation is impossible on nonparametric models, which have finite support. Our results illustrate the accuracy of this extrapolation.

Given a trained statistical model and an estimated p-value, we hypothesize that input sets of cavities exhibiting a high p-value will contain cavities with identical binding preferences, while input sets of cavities exhibiting an unusually small p-value will contain cavities with different binding preferences. We will test this hypothesis in our results.

3.3. Data Set Construction

Protein Families. The serine protease and enolase superfamilies were selected for this study based on the criteria that each superfamily exhibit three subfamilies with distinct binding preferences, and that variations in binding preferences are caused by well known structural mechanisms.

Serine proteases hydrolyze peptide bonds through the recognition of adjacent amino acids with specificity subsites numbered $S4, S3, \ldots S1, S1', S2', \ldots, S4'$. Each subsite preferentially binds one amino acid before or after the hydrolyzed bond between $S1$ and $S1'$. Cavities in our data set are derived from the $S1$ subsite, which binds aromatics in chymotrypsins [58], positively charged amino acids in trypsins [59], and small hydrophobics in elastases [60].

Proteins in the enolase superfamily catalyze a reaction that abstracts a proton from carbons adjacent to a carboxylic acid [61]. Opposite an N-terminal "capping domain" [62], the C-terminal domain forms a TIM-barrel, which provides a stable scaffold for amino acids that act as acid/base catalysts for several different reactions [61]. Cavities in our data set, on these amino acids, were classified into three subfamilies that facilitate the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate, in enolase [63], convert (R)-mandelate to and from (S)-mandelate [64], in mandelate racemase, and reciprocally cycloisomerize cis,cis-muconate to and from muconolactone, in muconate-lactonizing enzyme [61].

Serine Protease Superfamily:
Trypsins: 2f91, 1fn8, 2eek, 1h4w, 1bzx, 1aq7, 1ane, 1aks, 1trn, 1a0j
Chymotrypsins: 1eq9, 8gch
Elastases: 1elt, 1b0e

Enolase Superfamily:
Enolases: 1e9i, 1iyx, 1pdy, 2pa6, 3otr, 1te6
Mandelate Racemase: 1mdr, 2ox4
Muconate Lactonizing Enzyme: 2pgw, 2zad

Fig. 3. **PDB codes of structures used.**

Selection. The Protein DataBank (PDB - 6.21.2011) [65] contains 676 Serine proteases from chymotrypsin, trypsin, and elastase subfamilies and 66 enolase superfamily structures from enolase, mandelate racemase, and muconate cycloisomerase subfamilies. From each set, we removed mutant and partially ordered structures. Because enolases have open and closed conformations, all closed or partially closed structures were removed. Next, structures with greater than 90% sequence identity were removed, with preference for structures associated with publications, resulting in 14 serine protease and 10 enolase structures (Figure 3). Within these structures, ions, waters, and other non-protein atoms were removed. Since hydrogens were unavailable in all structures, all hydrogens were removed for uniformity. Atypical amino acids (e.g. selenomethionines) were not removed.

Alignment. Using Ska [16], an algorithm for aligning protein structures, all serine protease structures were aligned to bovine gamma-chymotrypsin (pdb code: 8gch), and all enolase superfamily structures to mandelate racemase from pseudomonas putida (pdb code: 1mdr). All structures in both superfamilies exhibit identical folds, causing the alignments of these proteins to another structure to cause little ultimate variation. This effect was observed earlier [11], where alignments recomputed for every possibility generated identical results. Following structural alignment, solid representations of binding cavities were generated using a method described earlier [11].

4. Experimental Results

4.1. Validating the Statistical Model

We considered multiple parametric models that would represent the degree of volumetric similarity between binding cavities with identical binding preferences. Testing these models on the trypsin and enolase subfamilies of the serine proteases and the enolase superfamily, respectively, we observed that the log-normal and gamma distributions most closely reflected the volumetric similarity measurements observed.

Figure 4 illustrates this point on the volumetric similarity between pairs of serine protease cavities as an example: In Figure 4B, the log transformed volumetric similarity sample data visibly follows a normal distribution. Futhermore, inspecting the quantile-quantile plots relating the data in Figure 4B to gamma, Weibull, Pareto, Generalized Extreme Value, and Log-Normal distributions (Figures 4C-F), it is clear that the log-normal and gamma plots are more linear than the others.

Similar observations were made when modeling the distribution of volumetric similarity between pairs of enolase cavities, as well as triplets and quadruplets of serine protease cavities, though in general, it appears that log-normal distributions followed the data more closely than the gamma distribution. Based on these observations, we use the log-normal distribution to estimate p-values.

4.2. Classifying Cavity Similarity

Multipart cross-validation was used to fully test the predictive accuracy of VASP-I. First, we computed the statistical significance of volumetric similarity among cavities having binding preferences identical to the training set. For both trypsins and enolases, we left out two cavities, while training VASP-I on the volumetric similarity of pairs of cavities from the remaining cavities. This is illustrated in the case of enolase, at the bottom of Figure 5A. We then evaluated the statistical significance of the volumetric similarity of the left out pair. This process was repeated until every pair of cavities had been left out once. Based on the conventional standard of significance, .05, volumetric similarity was statistically insignificant in 42 out of 45 trypsin validation runs, and 13 out of 15 enolase validation runs.

Since the trypsin set was larger than the enolase set, we also performed leave-3-out and leave-4-out cross validation in the same manner (leave-4-out cross validation is diagrammed in the top of Figure 5A).

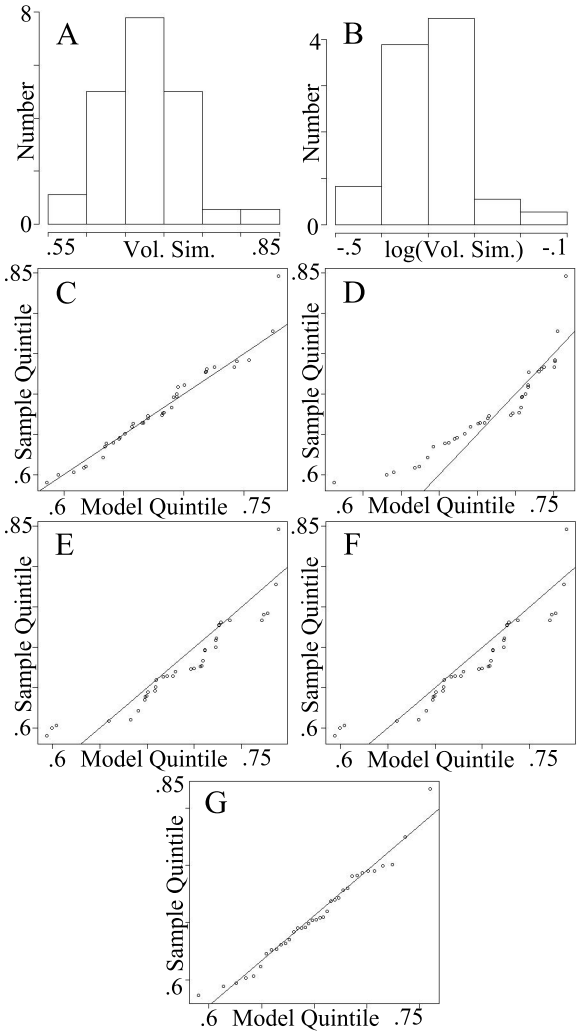

Fig. 4. **Distribution of volumetric similarity between pairs of trypsin cavities (A). Distribution of log transformed volumetric similarity values between pairs of trypsin cavities (B). Quantile-Quantile plots comparing the best fitting model against the log transformed sample data, for gamma (C), weibull (D), Pareto (E), Generalized Extreme Value (F), and Log-Normal (G) models.**

In leave-3-out, 106 out of 120 triplets had statistically insignificant volumetric similarity, and in leave-4-out, 170 out of 210 quadruplets had statistically insignificant volumetric similarity. The volumetric similarities of pairs, triplets and quadruplets of cavities were evenly distributed throughout the [0,1] range, and so were generally not statistically significant.

Next, we examined the ability of VASP-I to measure the statistical significance of volumetric similarity among cavities having binding preferences distinct from the training set. For both trypsins and enolases,

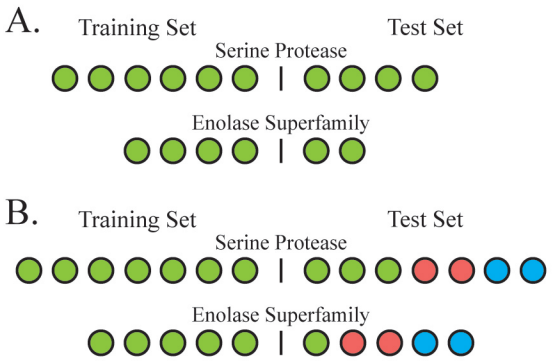

Fig. 5. **Circles represent individual cavities. Circles with different colors represent cavities with different binding preferences. (A) Diagrams illustrating how trypsins and enolases are separated, in one instance of a leave-4-out (top) and a leave-2-out (bottom) test intended to determine how effectively VASP-I classifies cavity sets with the same binding preferences. (B) Diagrams illustrating how serine proteases and enolase superfamily cavities are separated, in one instance of a leave-4-out (top) and a leave-2-out (bottom) test intended to determine how effective VASP-I classifies cavity sets with different binding preferences.**

we left out one cavity, and trained VASP-I on the volumetric similarity of the remaining pairs of trypsin or enolase cavities. Then, for the remaining trypsin or enolase cavity, we combined it in a training set with the other members of our dataset having different binding preferences. This configuration is illustrated, using enolases as an example, at the bottom of Figure 5B. In this case, 91 out of 100 serine protease pairs and 59 out of 60 enolase pairs were statistically significant.

Again, because of the larger size of the trypsin set, we performed leave-2-out and leave-3-out cross validation by training VASP-I on all but 2 and 3 trypsins, respectively. The remaining 2 (resp. 3) trypsins were combined with the other 4 non-trypsin serine proteases, enabling the generation of multiple sets of cavities with differing binding preferences. In leave-2-out validation we triplets of serine protease cavities and in leave-4-out validation we tested quadruplets, in order to ensure that no test triplet or quadruplet exhibited cavities with the same binding preferences. This configuration is illustrated at the top of Figure 5B. In leave-2-out cross validation, only 6 out of 900 sets with differing binding preferences were statistically insignificant, and in leave-3-out, only 9 out of 4200 were statistically insignificant. Pairs, triplets and quadruplets of cavities with heterogeneous binding preferences were almost always statistically significant.

5. Conclusions

We observed that sets of aligned cavities with differing binding preferences were almost always statistically significant, while sets of cavities with binding preferences identical to those used for training VASP-I were largely statistically insignificant. These results indicate that the statistical significance of volumetric similarity is an accurate predictive indicator of differences in binding preferences on our data set.

Given that VASP-I is an entirely geometric analysis of aligned binding cavities, the accuracy of its predictions is especially notable in the context of the broader range of biophysical effects that influence binding preferences here. A strong negative charge in the trypsin S1 pocket creates binding preferences for positively charged amino acids. The flexible capping domain in enolases also plays a considerable role in enolase specificity [62]. These other significant influences might be expected to create some degree of inaccuracy in our method, but it is apparent on our dataset that shape corresponded strongly with binding preference similarity. Nonetheless, the existence of these other biophysical influences points to logical directions for future work that incorporate additional biophysical information for more sophisticated predictions.

Another open question relates to the best fitting model distribution. On our small data set, we observed that the log-normal distribution appeared to fit the observed data more closely than gamma, Weibull, Pareto, and Generalized Extreme Value distributions, though the fit was only slightly better than the gamma distribution. Larger scale testing, planned for future work, may reveal a more closely fitting distribution that provides an accuracy improvement on already highly accurate estimations of statistical significance.

VASP-I represents the first application of statistical modeling to estimate the probability that a set of cavities exhibit identical binding preferences. As a result, VASP-I opens opportunities for new applications in the classification of cavities with differing binding preferences, even when it is unknown which cavities have different preferences. The comparative analysis of these cavities, once classified, may reveal new elements of protein structure that control specificity and ultimately point to new ways to design selective inhibitors.

Acknowledgment. The authors sincerely thank Viacheslav Y. Fofanov and Sean O'Keefe for critical discussions. This work was supported in part by start up funds from Lehigh University.

References

[1] A. C. Bishop, J. A. Ubersax, D. T. Petsch, D. P. Matheos, N. S. Gray, J. Blethrow, E. Shimizu, J. Z. Tsien, P. G. Schultz, M. D. Rose, J. L. Wood, D. O. Morgan, and K. M. Shokat, "A chemical switch for inhibitor-sensitive alleles of any protein kinase." *Nature*, vol. 407, no. 6802, pp. 395–401, Sep. 2000.

[2] Y. Liu, A. Bishop, L. Witucki, B. Kraybill, E. Shimizu, J. Tsien, J. Ubersax, J. Blethrow, D. O. Morgan, and K. M. Shokat, "Structural basis for selective inhibition of Src family kinases by PP1," *Chem Biol*, vol. 6, no. 9, pp. 671–678, 1999.

[3] L. Hedstrom, "Serine Protease Mechanism and Specificity," *Chem Rev*, vol. 102, no. 12, pp. 4501–24, Dec. 2002.

[4] S. D. Patel, C. Ciatto, C. P. Chen, F. Bahna, M. Rajebhosale, N. Arkus, I. Schieren, T. M. Jessell, B. Honig, S. R. Price, and L. Shapiro, "Type II cadherin ectodomain structures: implications for classical cadherin specificity." *Cell*, vol. 124, no. 6, pp. 1255–68, Mar. 2006.

[5] W. L. DeLano, "The PyMOL Molecular Graphics System," p. 0.98, 2002.

[6] D. Petrey and B. Honig, "GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences." *Method Enzymol*, vol. 374, no. 1991, pp. 492–509, Jan. 2003.

[7] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera–a visualization system for exploratory research and analysis." *J Comput Chem*, vol. 25, no. 13, pp. 1605–12, Oct. 2004.

[8] M. Sanches, S. Krauchenco, N. H. Martins, A. Gustchina, A. Wlodawer, and I. Polikarpov, "Structural characterization of b and non-b subtypes of hiv-protease: insights into the natural susceptibility to drug resistance development." *Journal of Molecular Biology*, vol. 369, no. 4, pp. 1029–1040, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17467738

[9] D. Petrey, Z. Xiang, C. L. Tang, L. Xie, M. Gimpelev, T. Mitros, C. S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I. Y. Koh, E. Alexov, and B. Honig, "Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling." *Proteins*, vol. 53 Suppl 6, no. February, pp. 430–435, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14579332

[10] H. Lee, Z. Li, A. Silkov, M. Fischer, D. Petrey, B. Honig, and D. Murray, "High-throughput computational structure-based characterization of protein families: Start domains and implications for structural genomics." *Journal of Structural and Functional Genomics*, vol. 11, no. 1, pp. 51–59, 2010.

[11] B. Y. Chen and B. Honig, "VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity," *PLoS Comput Biol*, vol. 6, no. 8, p. 11, 2010.

[12] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[13] R. Nussinov and H. J. Wolfson, "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques." *Proc Natl Acad Sci U S A*, vol. 88, no. 23, pp. 10 495–9, Dec. 1991.

[14] C. A. Orengo and W. R. Taylor, "SSAP: Sequential Structure Alignment Program for Protein Structure Comparison," *Method Enzymol*, vol. 266, pp. 617–635, 1996.

[15] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." *Protein Eng*, vol. 11, no. 9, pp. 739–47, Sep. 1998.

[16] 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.

[17] L. Holm and C. Sander, "Mapping the protein universe." *Science*, vol. 273, no. 5275, pp. 595–603, Aug. 1996.

[18] J. F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison." *Curr Opin Struct Biol*, vol. 6, no. 3, pp. 377–85, Jun. 1996.

[19] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments." *Proc Natl Acad Sci U S A*, vol. 105, no. 14, pp. 5441–6, Apr. 2008.

[20] Y. Ye and A. Godzik, "Flexible structure alignment by chaining aligned fragment pairs allowing twists." *Bioinformatics*, vol. 19 Suppl 2, no. 90002, pp. ii246–i255, 2003.

[21] M. Shatsky, R. Nussinov, and H. J. Wolfson, "FlexProt: alignment of flexible protein structures without a predefinition of hinge regions." *J Comput Biol*, vol. 11, no. 1, pp. 83–106, Jan. 2004.

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

[23] R. Kolodny, D. Petrey, and B. Honig, "Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction." *Curr Opin Struct Biol*, vol. 16, no. 3, pp. 393–8, Jun. 2006.

[24] C. a. Orengo, a. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH–a hierarchic classification of protein domain structures." *Structure*, vol. 5, no. 8, pp. 1093–108, Aug. 1997.

[25] S. S. Krishna and N. V. Grishin, "Structural drift: a possible path to protein fold change." *Bioinformatics*, vol. 21, no. 8, pp. 1308–10, Apr. 2005.

[26] D. Petrey and B. Honig, "Is protein classification necessary? Towards alternative approaches to function annotation," *Curr Opin Struct Biol*, vol. 19, no. 3, pp. 363–368, 2009.

[27] 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.

[28] J. A. Barker and J. M. Thornton, "An algorithm for constraint-based structural template matching : application to 3D templates with statistical analysis," *Bioinformatics*, vol. 19, no. 13, pp. 1644–1649, 2003.

[29] R. B. Russell, "Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution." *J Mol Biol*, vol. 279, no. 5, pp. 1211–27, Jun. 1998.

[30] B. J. Polacco and P. C. Babbitt, "Automated discovery of 3d motifs for protein function annotation." *Bioinformatics*, vol. 22, no. 6, pp. 723–730, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16410325

[31] 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, "Geometric Sieving : Automated Distributed Optimization of 3D Motifs for Protein Function Prediction," in *Proceedings of The Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006)*, 2006, pp. 500–515.

[32] B. Y. Chen, D. H. Bryant, A. E. Cruess, J. H. Bylund, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Composite motifs integrating multiple protein structures increase sensitivity for function prediction." *Comput Syst Bioinformatics Conf*, vol. 6, pp. 343–55, Jan. 2007.

[33] D. H. Bryant, M. Moll, B. Y. Chen, V. Y. Fofanov, and L. E. Kavraki, "Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction," *BMC Bioinformatics*, vol. 11, no. 242, 2010.

[34] 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. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21145898

[35] B. Y. Chen, D. H. Bryant, V. Y. Fofanov, D. M. Kristensen, A. E. Cruess, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Cavity-aware motifs reduce false positives in protein function prediction." in *Proceedings of the 2006 IEEE Computational Systems Bioinformatics Conference (CSB 2006)*, Jan. 2006, pp. 311–23.

[36] ——, "Cavity scaling: automated refinement of cavity-aware motifs in protein function prediction." *J Bioinform Comput Biol*, vol. 5, no. 2a, pp. 353–82, Apr. 2007.

[37] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility." *J Mol Biol*, vol. 55, no. 3, pp. 379–400, Feb. 1971.

[38] M. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *Science*, vol. 221, no. 4612, pp. 709–713, Aug. 1983.

[39] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov, "Molecular shape comparisons in searches for active sites and functional similarity." *Protein Eng*, vol. 11, no. 4, pp. 263–77, Apr. 1998.

[40] K. Kinoshita and H. Nakamura, "Identification of the ligand binding sites on the molecular surface of proteins," *Protein Sci*, vol. 14, pp. 711–718, 2005.

[41] R. A. Laskowski, "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions." *J Mol Graph*, vol. 13, no. 5, pp. 323–30, 307–8, Oct. 1995.

[42] T. A. Binkowski, "CASTp: Computed Atlas of Surface Topography of proteins," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3352–3355, Jul. 2003.

[43] T. A. Binkowski, L. Adamian, and J. Liang, "Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns," *J Mol Biol*, vol. 332, no. 2, pp. 505–526, Sep. 2003.

[44] T. A. Binkowski and A. Joachimiak, "Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites." *BMC Struct Biol*, vol. 8, p. 45, Jan. 2008.

[45] D. W. Ritchie and G. J. L. Kemp, "Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces," *J Comput Chem*, vol. 20, no. 4, p. 383, Mar. 1999.

[46] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors," in *European Symposium on Geometry Processing 2003*, 2003.

[47] A. Kahraman, R. J. Morris, R. a. Laskowski, and J. M. Thornton, "Shape variation in protein binding pockets and their ligands." *J Mol Biol*, vol. 368, no. 1, pp. 283–301, Apr. 2007.

[48] X. Zhang, C. L. Bajaj, B. Kwon, T. J. Dolinsky, J. E. Nielsen, and N. A. Baker, "Application of new multi-resolution methods for the comparison of biomolecular electrostatic properties in the absense of global structural similarity," *Multiscale Model Simul*, vol. 5, no. 4, pp. 1196–1213, 2006.

[49] M. Nayal and B. Honig, "On the Nature of Cavities on Protein Surfaces : Application to the Identification of Drug-Binding Sites," *Proteins: Struct. Funct. Genet.*, vol. 63, pp. 892–906, 2006.

[50] F. Glaser, R. J. Morris, R. J. Najmanovich, R. a. Laskowski, and J. M. Thornton, "A method for localizing ligand binding pockets in protein structures." *Proteins: Struct. Funct. Bioinf.*, vol. 62, no. 2, pp. 479–88, Feb. 2006.

[51] R. G. Coleman and K. A. Sharp, "Travel depth, a new shape descriptor for macromolecules: application to ligand binding." *J Mol Biol*, vol. 362, no. 3, pp. 441–58, Sep. 2006.

[52] A. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces." *J Mol Biol*, vol. 280, no. 1, pp. 1–9, Jul. 1998.

[53] A. Stark, S. Sunyaev, and R. B. Russell, "A Model for Statistical Significance of Local Similarities in Structure," *J Mol Biol*, vol. 326, pp. 1307–1316, 2003.

[54] C. P. Chen, S. Posy, A. Ben-Shaul, L. Shapiro, and B. Honig, "Specificity of cell-cell adhesion by classical cadherins: Critical role for low-affinity dimerization through beta-strand swapping." *Proc Natl Acad Sci U S A*, vol. 102, no. 24, pp. 8531–6, Jun. 2005.

[55] B. Chen and S. Bandyopadhyay, "VASP-S: A Volumetric Analysis and Statistical Model for Predicting Steric Influences on Protein-Ligand Binding Specificity," in *Proceedings of 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2011, p. accepted.

[56] H. Voelcker and A. Requicha, "Geometric Modeling of Mechanical Parts and Processes," *Computer*, vol. 10, no. 12, pp. 48–57, Dec. 1977.

[57] J. Schaer and M. Stone, "Face traverses and a volume algorithm for polyhedra," *Lect Notes Comput Sc*, vol. 555/1991, pp. 290–297, 1991.

[58] K. Morihara and H. Tsuzuki, "Comparison of the specificities of various serine proteinases from microorganisms," *Arch Biochem Biophys*, vol. 129, no. 2, pp. 620–634, 1969.

[59] L. Gráf, a. Jancsó, 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." *Proc Natl Acad Sci U S A*, vol. 85, no. 14, pp. 4961–5, Jul. 1988.

[60] G. I. Berglund, A. O. Smalas, H. Outzen, and N. P. Willassen, "Purification and characterization of pancreatic elastase from North Atlantic salmon (Salmo salar)." *Mol Mar Biol Biotechnol*, vol. 7, no. 2, pp. 105–14, Jun. 1998.

[61] 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 alpha-protons of carboxylic acids." *Biochemistry*, vol. 35, no. 51, pp. 16 489–501, Dec. 1996.

[62] J. F. Rakus, A. A. Fedorov, E. V. Fedorov, M. E. Glasner, B. K. Hubbard, J. D. Delli, P. C. Babbitt, S. C. Almo, and J. A. Gerlt, "Evolution of enzymatic activities in the enolase superfamily: L-rhamnonate dehydratase." *Biochemistry*, vol. 47, no. 38, pp. 9944–54, Sep. 2008.

[63] K. Kühnel and B. F. Luisi, "Crystal structure of the Escherichia coli RNA degradosome component enolase." *J Mol Biol*, vol. 313, no. 3, pp. 583–92, Oct. 2001.

[64] 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. 5662–9, May 1996.

[65] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank." *Nucleic Acids Res*, vol. 28, no. 1, pp. 235–42, Jan. 2000.