VASP-S: A Volumetric Analysis and Statistical Model for Predicting Steric Influences on Protein-Ligand Binding Specificity

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

*Many fields seek to identify steric influences in protein-ligand binding specificity. In some cases, these influences can be found by visually comparing protein structures, but subtler influences, whose significance may only be apparent from the analysis of many structures, are harder to find. To assist this process, we present VASP-S (*V*olumetric* A*nalysis of* S*urface* P*roperties with* S*tatistics), an unsupervised volumetric analysis and statistical model for isolating statistically significant structural variations that may influence specificity. We applied these methods to analyze sequentially nonredundant structural representatives of two well-studied protein families: the canonical serine proteases and the enolase superfamily. We observed that statistically significant structural variations, as identified by VASP-S, reproduced experimentally established determinants of specificity. These results suggest that unsupervised methods, supported by statistical models, may be able to automatically identify variations that sterically influence specific binding in catalytic sites.*

1. Introduction

A shared challenge in structural biology, protein engineering, and drug design is the elucidation of the molecular mechanisms underpinning protein-ligand binding specificity. Understanding these mechanisms may reveal how protein selectivity organizes crowded molecular environments [1], how proteins could be mutated to alter binding preferences [2], or how elements of protein structures affect drug resistance [3]. The heart of this challenge lies in the fact that the molecular mechanism that drives preferential binding arises from multiple structural elements, such as amino acid sidechains [3] and cavity regions [4]. When the elements involved in this mechanism are unknown, a combinatorial space of possibilities must be ruled out in order to determine the actual causes of preferential binding.

Undertaking this process experimentally can be impractical without a guiding hypothesis. One source of such hypotheses lies in the differing shapes of ligand binding cavities in families of closely related proteins. Among aligned protein cavities, regions where cavities vary may cause differing substrates to bind. Overlapping regions might bind a molecular fragment that is common to substrates acted on by the entire family. This effect occurs in serine proteases, where binding cavities differ in size to better accommodate differently sized substrates [5]. A similar effect can be seen in the enolase superfamily, where different amino acids arranged around a common scaffold enable different reactions to be catalyzed [6]–[8]. Variations of this kind can sometimes be identified by visual inspection, but when many variations exist, it is frequently unclear which of the variations found, if any, are significant enough to evaluate experimentally as potential specificity determinants.

This paper reports a computational method for identifying significant cavity variations called VASP-S (**V**olumetric **A**nalysis of **S**urface **P**roperties with **S**tatistics). Applying existing methods [9] to represent protein structures and cavities as geometric solids, and compare them with Boolean Set operations (Figure 1), VASP-S adds the new capability to separate individual regions, called *fragments*, that lie within one cavity and not within another (e.g. Figure 1h,i). Thus, a fragment is one of potentially several variations between two cavities. VASP-S can also construct a statistical model of the volumes of fragments that occur in the binding cavities of proteins with identical specificity, enabling it to detect fragments with statistically significant (e.g. unusually large) volume. We hypothesize that the ex-

1. Corresponding Author

istence of a statistically significant fragment indicates a variation in cavity shape that is large enough to accommodate different ligands, and thus a cause of different binding preferences.

We tested this hypothesis on two sequentially nonredundant families of protein structures: the serine protease and enolase superfamilies. On these data sets, in cross-validated experimentation, we observed that the largest fragments between cavities with different binding preferences were statistically significant. In many cases, they reproduced experimentally established influences on specificity. We also observed that the volumes of fragments between cavities with identical binding preferences were almost always statistically insignificant. These results point to applications for discovering the structural mechanisms that affect binding preferences, or gathering evidence that such mechanisms do not exist.

2. Related Work

The solid representations of protein structures and binding cavities used by VASP-S differ considerably from typical comparison algorithms, which use point-based and surface-based representations. Point-based representations encode atoms in protein structures using points in three dimensions [10]–[14], matrices of distances between points [15], and nodes in geometric graphs [16], [17]. These representations are traditionally applied to rigidly superpose and align whole protein structures, but, more recently, flexible methods [18] have also emerged. A second type of point-based representation is specialized for the comparison of functional sites, using motifs in three dimensions that encode atoms in catalytic sites [19]–[21], evolutionarily significant amino acids [22], "pseudocenters" representing protein-ligand interactions [23], and pseudoatoms representing amino acid sidechains [24]. Point-based methods exhibit extreme efficiency, enabling them to rapidly search for evolutionarily remote homologs [19], [20], [25] in large databases of protein structure [26], but they are not intended for isolating variations in empty cavity regions, like the methods presented here.

Surface-based representations employ closed surfaces or surface patches to represent or approximate solvent-accessible shape [27], [28]. These representations are built from triangular meshes [29], [30], alpha shapes [31]–[33], three dimensional grids [34], and spherical harmonics [35]–[37]. In some cases surface representations have been applied for the comparison of protein structures [29], [30] and electrostatic potentials [38], as well as in hybrid representations that

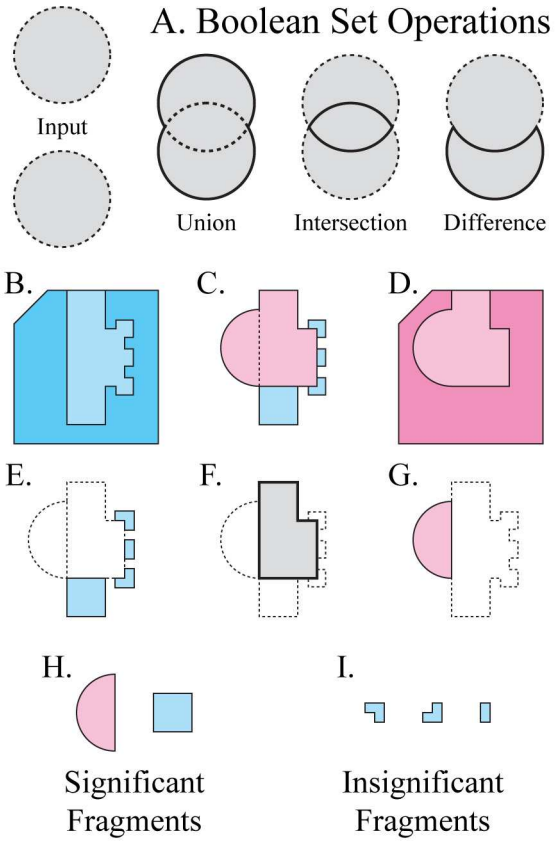

Fig. 1. **Isolating Significant Cavity Regions with Boolean Set Operations. A. A diagram of Boolean Set operations, showing the borders of input regions (dotted) and output (solid), in grey. B,D) Polygons representing regions occupied by protein X (blue) and protein Y (red), their molecular surfaces (black lines), and their binding cavities x (light blue) and y (light red). C) Superimposition of x and y, based on a whole structure alignment of X and Y. In E, F, and G, the superposition of x and y is depicted as dotted lines. Regions in solid lines were computed with Boolean set operations. E) The difference of x and y. F) The intersection of x and y. G) The difference of y and x. H) Statistically significant fragments from E and G. I) Statistically insignificant fragments from E and G.**

combine point-based and surface-based information [39], but they have had widest application in the identification of cavities and hot spots [40] in protein surfaces [31], [41]–[43]. While surface-based methods identify and compare surface cavities, VASP-S offers the new capability of isolating individual variations within cavities.

To the space of point-based and surface-based representations, the geometric solids used in VASP-S [9] contribute an orthogonal third representation for the comparison of protein structures. Comparisons with Boolean set operations are related to volumetric methods that measure volume differences in catalytic

sites [33], [44] and electrostatic isocontours [38], and to methods for identifying regions where substrates overlap [45], but VASP-S differs because it can isolate individual varying regions, as we illustrate in Section 3.1. In our results, this capability enables steric influences on ligand binding specificity to be identified automatically.

Statistical modeling plays a critical role in the unsupervised comparison of protein structures, especially in the identification of geometrically similar catalytic sites at remote evolutionary distances. In that application, statistical modeling enables the computation of data-specific thresholds to identify catalytic sites that are improbably similar, and thus potential markers of functional similarity. Several independent results, using Gaussian mixture models [20], extreme value distributions [46], nonparametric models [19], and empirical models [32], [47], have observed that statistically significant geometric similarity is an accurate marker of similar functional sites. In this work, VASP-S introduces a new application for statistical modeling by characterizing the statistical significance of variations in catalytic sites.

3. Methods

As described in earlier work [9], beginning with solid representations of two aligned cavities, Boolean set operations can identify regions within one cavity and not another, but they do not separate individual fragments. This section describes, first, a general approach for the topological separation of fragments following Boolean set operations, second, how this information can be used to train a statistical model of fragment volume, and finally, the construction of data sets to test these methods.

3.1. Identifying Fragments

We begin with two geometric solids, A (Figure 2a) and B (Figure 2c), representing cavities from aligned protein structures. Using Boolean Set operations, we compute AB, the region inside A and not inside B, as well as BA, the region inside B and not A. Like A and B, AB and BA are geometric solids represented by closed triangular meshes. Fragments from AB and BA are, together, referred to as the set of fragments relating to cavities A and B (Figure 2b).

First, we translate the triangular mesh of AB and BA into a graph G, mapping corners to graph nodes, and triangle edges to graph edges. Since G is likely to have several connected components, we separate each connected component into an individual graph

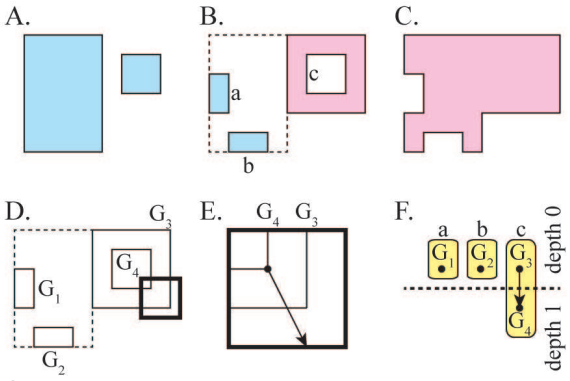

Fig. 2. **Input cavity A (light blue,A) and B (light red,C). Fragments derived from AB and BA, labeled with lower case letters, original cavities outlined in dotted lines (B). Fragments translated into connected components (solid lines, labeled G_1-G_4), original cavities outlined in dotted lines (D). A ray (arrow) from a corner of G_4, drawn from the enlarged perspective of the dark box in D (E). The directed graph, H, with nodes $G_1 - G_4$, connected based on containment. The dotted line separates nodes of different depths. Yellow ovals indicate subgraphs relating to the same fragment, labeled with lower case letters (F). Cavity A is shown atypically, in a disconnected manner, to illustrate the algorithm.**

G_i (Figure 2d). This can be accomplished through depth first search in linear time [48]. Each connected component does not necessarily represent an individual fragment, because fragments occasionally contain interior voids, as illustrated in Figure 2b, that are composed of multiple disconnected components.

Next, we determine which connected components reside within another connected component. This is accomplished with ray casting (Figure 2e). For each component G_i, a ray, beginning at one point on G_i is pointed in a random direction, and the number of intersections with other components is counted. If the ray intersects another component G_j an even number of times, then we say that G_i is not inside G_j. Alternately, if the ray intersects G_j an odd number of times, then we say that G_i is inside G_j. For all pairs G_i and G_j, we determine which contains the other.

Finally, we represent the pattern of containment as a directed acyclic graph H (Figure 2f), where each node represents a graph G_i, and an edge from G_i to G_j indicates that G_i is within G_j according to the test above. Using the topology of H, we can determine which G_i are part of the same fragment: First, we identify subgraphs G_{i_k} that are not contained inside any other graph, because their in-degree is zero. From each G_{i_k}, we perform a depth first search and assign an integer depth d to each G_i considered. Since H is an acyclic graph, some G_i may be visited more than once. In these cases, if the number of edges traversed

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

Fig. 3. **Estimating the probability p of observing a fragment F', with volume $v(F') \geq v(F)$, using the mean ($\overline{x}$) and variance (s) of the distribution of the log-transformed sample values to estimate μ and σ. Φ is the cumulative distribution function of the standard normal distribution**

from the originating G_{i_j} to the current G_i is greater than the depth assigned already, d is reassigned the larger value. This reassignment process determines the number of times each graph G_i is nested within the entire group of connected components, since the largest possible depth reflects the actual number of times one graph is nested inside the others.

Finally, we separate H into subgraphs. Each G_i with an even depth d is an exterior surface for one fragment. Based on the topology of H, each G_i with an odd depth d resides inside an exterior surface with even depth equal to $d-1$. Thus, we can associate the graphs G_i into groups that are all part of the same fragment, and output the fragment. This correctly separates fragments of arbitrary nesting.

3.2. A Statistical Model of Fragment Volume

Our statistical model is based on a hypothesis testing framework that detects fragments with volume large enough to be statistically significant, i.e. unlikely to occur by random chance. The underlying assumption of our model is that fragments derived from cavities with no difference in specificity will have *small* volumes related to incidental and functionally irrelevant structural variation. Alternatively, if there exists a structural variation in one cavity large enough to create a steric influence on specificity, then the fragment generated by the variation between the cavities will have *unusually large* volume. Thus, for a query fragment F, based on cavities A and B, our null hypothesis asserts that the volume of F, $v(F)$, is *small*. The alternative hypothesis asserts that $v(F)$ is *unusually large*. Since they are logical complements, exactly one of these hypotheses can hold for any fragment F.

We test the null hypothesis by first assuming that it holds for F, and then estimating the probability p of randomly observing another fragment F', with volume $v(F') \geq v(F)$. If the probability of randomly observing another fragment with larger volume is improbably low, typically below 0.05, then it is hard to continue assuming that F is small. In this circumstance, the null hypothesis is rejected as improbable, leaving us to favor the alternative hypothesis, that F is large. The biological interpretation of this decision follows from our underlying assumption: F is unusually large, and may thus be a structural variation in either A or B that creates a steric influence on specificity. This statement is a prediction based on quantified evidence, not a statement of fact.

In order to perform this prediction, we must estimate the probability p, which requires us to first train the statistical model. Training begins with aligned cavities from the training sets described in Section 3.3. First, we separate the fragments generated between all pairs of cavities using the method described in Section 3.1. Using the Surveyor's Formula [49], which provides a rapid and very accurate estimation of volume in a closed surface, we compute the volume of each fragment. These data are represented in a frequency distribution D (See Figure 5A). The shape of D closely fits a log-normal distribution, as seen in Section 4.1.

Since D fits $log\text{-}normal(\mu, \sigma)$, we can use the log-normal distribution to smoothly estimate the probability p of observing any a fragment F', with volume greater than or equal to the volume of our query fragment $v(F)$. This estimation occurs when we realize that the mean μ and the variance σ of the log-normal distribution are unknown: we estimate μ and σ with the mean $\overline{x}$ and variance s from the distribution of log-transformed values of D. We can thus estimate p using Equation 1. p is the proportion of the volume under the log-normal curve to the right of $v(F)$, relative to the total volume under the curve ($x \geq 0$).

Fitting the log-normal function to D enables this probability to be estimated without the discretizing effect of the training data. Also, assuming that the log-normal distribution is a sufficiently accurate estimation of the underlying probability density function, we can extrapolate the probability beyond the largest volume observed in our training data. Such extrapolation would not be possible using nonparametric models, which have finite support. The accuracy of this extrapolation is illustrated in our results.

Having trained our statistical model on fragments derived from cavities with identical binding specificities, we hypothesize that our statistical model will behave as follows: Fragments generated between a cavity with similar binding preferences and a cavity with different binding preferences are be expected to have a statistically significant fragment, if there exists a steric influence on specificity. Likewise, for two cavities having the same binding preferences as the training set, fragments generated between them are not expected to be statistically significant. We test this hypothesis in our experimental results.

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. 4. **PDB codes of structures used.**

3.3. Data Set Construction

Protein Families. The serine protease and the enolase superfamilies were selected for demonstrating VASP-S because several sequentially nonredundant structures exist for both superfamilies. Each superfamily contained at least three subfamilies with distinct binding preferences and at least two nonredundant structural representatives in each subfamily.

Serine proteases catalyze the hydrolysis of specific peptide bonds by recognizing neighboring 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 $S1'$. Our demonstration, on three subfamilies, focuses on the $S1$ subsite, which binds aromatics in chymotrypsins [50], positively charged amino acids in trypsins [51], and small hydrophobics in elastases [52].

Members of the enolase superfamily catalyze a variety of reactions that involve the abstraction of a proton from a carbon adjacent to a carboxylic acid [6]. Assisted by an N-terminal "capping domain" [53], amino acids at the C-terminal ends of beta sheets in a conserved TIM-barrel act as acid/base catalysts to facilitate several different reactions [6]. Our demonstration, on three subfamilies, is focused on the primary catalytic site, which facilitates the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate in enolase, [54], the conversion of (R)-mandelate to and from (S)-mandelate [55] in mandelate racemase, and reciprocal cycloisomerization of cis,cis-muconate and muconolactone in muconate-lactonizing enzyme [6].

Selection. The Protein DataBank (PDB - 6.21.2011) [26] contains 676 Serine proteases from chymotrypsin, trypsin, and elastase subfamilies and 66 enolase superfamily structures from enolase, mandelate racemase, and muconate cycloisomerase subfamiles. 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

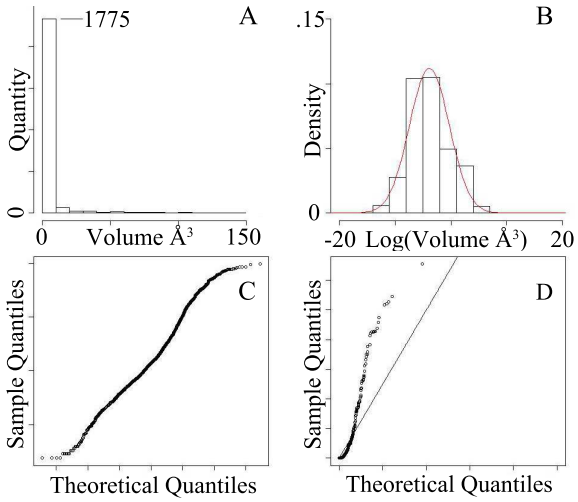

Fig. 5. **Histogram counting of the number of fragments between pairs of trypsin cavities, in volume bins (A), and in log(Volume) bins plotted against the best fitting Gaussian (B). Quantile-quantile plots of trypsin fragment volumes relative to the best fitting Gaussian (C) and Gamma distributions (D).**

14 serine protease and 10 enolase structures (Figure 4). 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. Ska [14], an algorithm for whole-protein structure alignment, was used to align all serine protease structures to bovine gamma-chymotrypsin (pdb code: 8gch), and all enolase superfamily structures to mandelate racemase from pseudomonas putida (pdb code: 1mdr). Since proteins in these datasets have identical folds, alignments to a different structure has little effect: This was observed earlier [9], where cavity comparisons, recomputed with the same method, generated identical results. Beginning with this alignment, solid geometric representations of binding cavities were generated with a method described earlier [9], based on cavities defined in SCREEN [41].

4. Experimental Results

4.1. Validating the Statistical Model

We constructed a statistical model based on all trypsin cavities and a second based on all enolase cavities. The distribution of fragment volumes between trypsin and enolase cavities is illustrated in Figure 5a. Fragments with volumes near zero dominated, though both distributions exhibited a positive tail. Seeking the best fitting parametric model, we tested gamma,

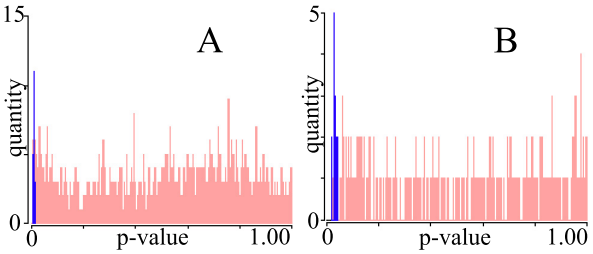

Fig. 6. **Histograms of fragments in p-value bins, depicting fragments from cavities with different specificities (dark blue) and fragments from cavities with similar specificities (light red), in proteins from serine protease (A) and enolase (B) superfamilies.**

Weibull, Pareto, generalized extreme value, and log-normal distributions.

In both Trypsin and Enolase sets, we observed that log-normal distributions fit the observed data best. This is apparent in part when considering how well the Gaussian distribution fits the log of the fragment volumes (Figure 5b), but even more so when considering the quantile-quantile (q-q) plots comparing the log of observed fragment volumes versus a Gaussian distribution (Figure 5c). Other distributions considered led to poorer q-q plots: The second best, in both cases was the gamma distribution (Figure 5d). Enolase plots (not shown) were similar in overall shape, and supported the same conclusions. All plots are available here: www.cse.lehigh.edu/∼chen/papers/BIBM2011

4.2. Calculating Fragment Significance

We calculated the statistical significance of fragments that exist between cavities with different binding specificities, in a leave-one-out manner: First, the statistical model was trained on all but one trypsin or enolase cavity. With the remaining trypsin or enolase cavity and each of the non-trypsin or non-enolase cavities, we determined the largest fragment, and measured its p-value. This process was repeated once each trypsin and enolase cavity, producing 40 trypsin fragments and 36 enolase fragments (Figure 6, dark blue).

We also calculated the statistical significance of fragments that exist between cavities with the same binding specificities. In a leave-two-out experiment, we first trained the statistical model with all but two trypsins or elastases. With the remaining two trypsins or elastases we computed the p-value of all fragments. This process was repeated for every combination of two members in the trypsin and enolase sets, producing 1893 trypsin fragments and 340 enolase fragments (Figure 6, light red).

The largest fragments from cavities with different binding preferences were always statistically significant, following the standard 0.05 threshold of statistical significance. By the same standard, fragments from cavities with identical binding preferences were rarely significant, exhibiting widely distributed p-values.

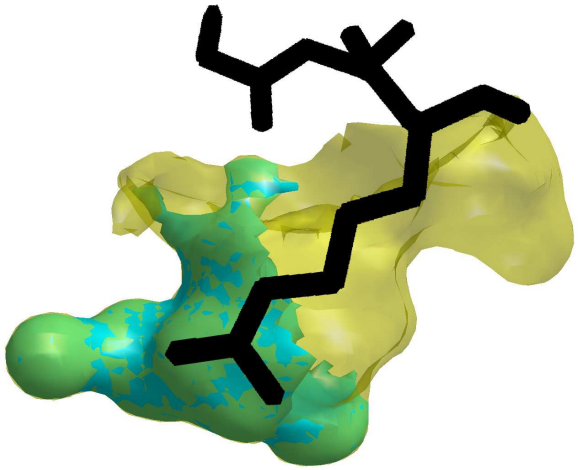

Fig. 7. **S1 specificity site of Atlantic salmon trypsin (transparent, yellow). Statistically significant fragment within the trypsin cavity and not within the S1 specificity site of porcine pancreatic elastase (opaque, teal). Gly-Ala-Arg peptide modeled from Fusarium oxysporum trypsin (black sticks).**

4.3. Verifying Fragment Function

Statistically significant fragments identified several variations in cavity shape that influence binding preferences. One example, illustrated in Figure 7 depicts a statistically significant fragment that is within the S1 specificity site of Atlantic salmon trypsin (pdb reference: 1a0j) and not within the S1 specificity site of porcine pancreatic elastase (pdb reference: 1b0e). The fragment occupies a volume of 144 Å^3, and is the largest of several differences between these cavities. The position of the fragment highlights a region in the trypsin cavity that extends deeper than the elastase cavity. This region is essential for accommodating the longer, positively charged substrates preferred by trypsins [51]. A modeled Gly-Ala-Arg peptide illustrates this point in Figure 7. Much like this example, similar significant fragments could be found between all trypsin and elastase cavities, as well as between trypsin and chymotrypsin cavities. A second class of related effects were observed in enolase cavities, where sidechains protruding from different parts of the conserved beta-barrel scaffold created differences in cavity shape that accommodate different catalytic reactions.

5. Conclusions

We have presented a computational method for generating fragments and a statistical model for estimating the significance of fragment volume. To our knowledge, VASP-S is the first algorithm capable of separating individual differences in cavity shape, and also the first to measure their statistical significance, creating a new strategy for identifying influences on protein-ligand binding specificity.

After verifying the choice of distributions for our statistical model, we demonstrated VASP-S by identifying statistically significant fragments by volume, from serine protease and enolase cavities. In both cases, the largest fragment between cavities with different binding preferences was always statistically significant, while all fragments between cavities with identical binding preferences were rarely so. By identifying differences in binding cavity shape that are too large to have randomly occurred between cavities with identical binding preferences, this approach predicts cavity regions that drive different binding preferences.

We verified the accuracy of some of these predictions by relating them to experimentally established observations, where possible. In both serine protease and enolase datasets, the most statistically significant fragment was frequently a difference in binding cavity geometry that enabled the accommodation of differently shaped substrates. While other physical phenomena (e.g. electrostatics [51]) are known to influence specificity in both datasets, statistically significant fragments remained strong markers of structural influences on specificity. On other data sets, variations in shape may not be as strongly correlated with specificity. This possibility points to potentials for future work.

Applications of VASP-S exist in contexts where the detailed comparison of protein cavities is required. For example, in drug design, statistically significant fragments in ligand binding cavities may identify variations in cavity shape that can be exploited for more selective inhibitors. In concert with other sources of information, technologies like VASP-S offers new tools for the elucidation of of protein-ligand interactions.

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

References

[1] 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.

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

[3] 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.

[4] R. A. Musah, G. M. Jensen, S. W. Bunte, R. J. Rosenfeld, and D. B. Goodin, "Artificial protein cavities as specific ligand-binding templates: characterization of an engineered heterocyclic cation-binding site that preserves the evolved specificity of the parent protein." *J Mol Biol*, vol. 315, no. 4, pp. 845–57, Jan. 2002.

[5] D. Shotton and H. Watson, "The three-dimensional structure of porcine pancreatic elastase," *Philos Trans R Soc London [Biol]*, vol. 257, no. 813, pp. 111–118, 1970.

[6] 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.

[7] M. S. Hasson, I. Schlichting, J. Moulai, K. Taylor, W. C. Barrett, G. L. Kenyon, P. C. Babbitt, J. A. Gerlt, G. A. Petsko, and D. Ringe, "Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase." *Proc Natl Acad Sci U S A*, vol. 95, no. 18, pp. 10 396–401, Sep. 1998.

[8] J. A. Gerlt and P. C. Babbitt, "Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies." *Annu Rev Biochem*, vol. 70, pp. 209–46, Jan. 2001.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] B. Y. Chen, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Algorithms for structural comparison and statistical analysis of 3D protein motifs." *Pac Symp Biocomput*, vol. 345, pp. 334–45, Jan. 2005.

[23] M. Shatsky, A. Shulman-peleg, R. Nussinov, and H. J, "Recognition of Binding Patterns Common to a Set of Protein Structures," *Lect Notes Comput Sc*, vol. 3500, pp. 440–455, 2005.

[24] S. Schmitt, D. Kuhn, and G. Klebe, "A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology," *J Mol Biol*, vol. 323, no. 2, pp. 387–406, Oct. 2002.

[25] M. Moll, D. H. Bryant, and L. E. Kavraki, "The LabelHash algorithm for substructure matching." *BMC Bioinformatics*, vol. 11, no. 1, p. 555, Jan. 2010.

[26] 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.

[27] 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.

[28] M. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *Science (80-)*, vol. 221, no. 4612, pp. 709–713, Aug. 1983.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] I. Wallach and R. H. Lilien, "Prediction of sub-cavity binding preferences using an adaptive physicochemical structure representation." *Bioinformatics*, vol. 25, no. 12, pp. i296–304, Jun. 2009.

[45] M. N. L. Nalam, A. Ali, M. D. Altman, G. S. K. K. Reddy, H. Cao, M. K. Gilson, B. Tidor, T. M. Rana, and C. A. Schiffer, "Evaluating the Substrate-Envelope Hypothesis : Structural Analysis of Novel HIV-1 Protease Inhibitors Designed To Be Robust against Drug Resistance," *J Virol*, vol. 84, no. 10, pp. 5368–5378, 2010.

[46] 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.

[47] 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.

[48] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, "Introduction to Algorithms , Second Edition," *Computer*, vol. 7, no. 9, p. 984, 2001.

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

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.