A Statistical Model of Electrostatic Isopotential Variation in Serine Protease Binding Cavities

Rachel Y. Okun
Dept. of Computer Science and Engineering
Lehigh University
Bethlehem, PA, USA
ryo218@lehigh.edu

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

***Abstract*—This paper presents EPAC (Electrostatic isoPotential Analytical Comparative model), the first statistical model for evaluating the geometric similarity of electrostatic fields. Beginning with aligned binding cavities, EPAC measures similarity based on the overlapping volume of isopotentials inside ligand binding cavities. We tested the accuracy of our model on two subfamilies of the serine protease superfamily, demonstrating that EPAC effectively identifies binding sites that prefer differently charged substrates. For example, EPAC identified subtle electrostatic variations in proteins that might be expected to be more similar, such as the difference between typical trypsins and a trypsin with a phosphorylated tyrosine nearby the binding site. These results point to applications in the unsupervised comparison of many binding sites from a purely electrostatic perspective, in the search of subtle electrostatic variations that could influence binding specificity.**

I. Introduction

Statistical models of geometric variation are widely used in protein structure comparison. These models are trained to establish the typical degree of geometric variation that occurs between protein structures with similar function [1]–[4] or specificity [5]–[8]. Once trained, effective models can automatically detect proteins with atypical structural similarity with constant computational cost [9]–[11]. As a result, statistical models enable investigators to identify unusual proteins at a large scale without human supervision.

Statistical models are closely tied to geometric measures of similarity. Most methods reported to date model the variation of root mean squared distance (RMSD) between the backbone atoms of whole protein structures [12]–[18]. A second category of methods model variation between atoms representing protein ligand binding sites [9], [19]–[22]. Naturally, many biophysical variations exist between proteins that are not represented by RMSD, such as the presence of cavities [23], [24], the number of atoms being compared [25], evolutionary significance [3] or electrostatics [20], [26], [27]. These variations can influence the model because of their effect on membership in the training set, but their influence is contingent on detection of geometric similarity.

This paper examines the effect of modeling geometric similarity in electrostatic isopotentials (EP) between ligand binding cavities. Beginning with two aligned binding cavities, we measure similarity between electrostatic isopotentials based on their degree of volumetric overlap. We use this value to train a parametric statistical model of the similarities that exist between binding cavities with identical binding preferences. The EP similarity observed from subsequent comparisons can then be evaluated according to the model, generating a p-value that describes how unusual it is relative to the training set.

In our results, we demonstrate on two families of serine proteases that our parametric model distinguishes pairs of proteins with similar electrostatic binding preferences from those with different electrostatic binding preferences. This unique approach, the first statistical model of purely electrostatic similarity, is totally independent of atomic comparison and it illustrates that alternatives to statistical models of RMSD are possible and that they can be effective tools for specificity annotation. In addition to applications in specificity annotation, where it could be used for finding electrostatic variations that likely cause differences in specificity, EPAC could also be applied in concert with existing statistical models of structural variation. By integrating probability estimates from multiple sources, a compound statistical model could point to the potential for influences on specificity from multiple biophysical sources.

II. Methods

This section describes, first, how we compute EP similarity, how we train to the statistical model, how statistical significance is determined, and how the data set is constructed.

A. *Measuring Electrostatic Similarities*

We begin with two aligned protein structures and a potential threshold k, which determines the potential at which isopotential geometry in the binding sites of the proteins will be compared. We evaluate electrostatic similarity using Boolean operations from constructive solid geometry (CSG, Figure 1). Beginning with two aligned protein structures, a bound ligand, and a threshold of electrostatic potential, we generate a volumetric description of each binding cavity using a subcavity method described earlier [28] (Figure 1d,e). Next, we use CSG to compute the volumetric intersection of the two cavities. This intersecting region, is solvent accessible in both cavities, and thus a region where steric differences do not interfere

[*] Corresponding author.

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

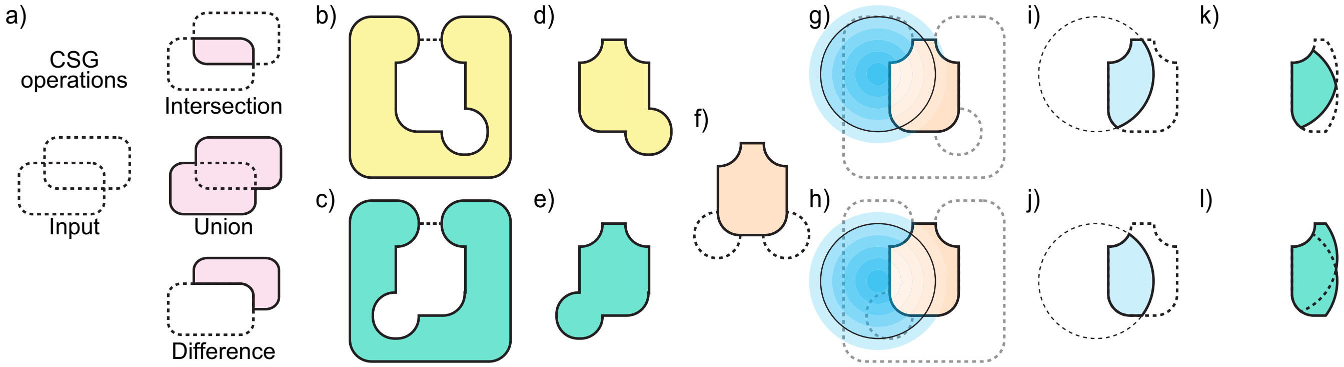

Fig. 1. Comparing cavity fields. a) Boolean operations. Dotted lines define input regions. Solid lines define output regions. b,c) Molecular surface of two aligned proteins. Dotted lines denote boundaries of the binding cavities. d,e) Binding cavities. f) Intersection of binding cavities based on the alignment of the proteins (orange). g,h) Electrostatic potential fields illustrated as multiple positive isopotentials (transparent blue) over the conserved binding region (orange). The selected isopotential is highlighted with a thin black line. i,j) Cavity fields of both proteins. k,l) The intersection (k, green with solid lines) and union (l, green with solid lines) of the cavity fields (dotted).

$$p(C_i') = 1 - \Phi(\frac{\log{(C_i')} - \mu}{\sigma}) \approx 1 - \Phi(\frac{\log{(C_i')} - x}{s}). \tag{1}$$

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

with electrostatic similarities or differences. We perform our electrostatic comparisons in this region.

Next, we use Delphi [29] to compute the electrostatic fields of both proteins, and VASP-E [30] to generate the electrostatic isopotential based on the provided potential threshold. The CSG intersection between the isopotential and the cavity intersection above yields a description of the electrostatic field of each protein within the region of the binding site that is shared by both proteins. For any given protein, we refer to this region as a *cavity field* defined on k.

Given a pair of aligned cavity fields c_0 and c_1, we define volumetric similarity $d(c_0, c_1)$ using the Jaccard index $d(C)$: [31]

$$d(C) = 1 - \frac{v(c_0 \cap c_1)}{v(c_0 \cup c_1)}$$

.

Here, $v(x)$ denotes the volume of space within some geometric solid x. We compute $v(x)$ with the Surveyor's Formula [31]. We subtract the fraction from 1, to produce an *electrostatic distance*: Cavity fields with substantial EP similarity yield distances close to zero, while the distance between very different cavity fields approaches one.

B. Statistical Model of Electrostatic Similarity

Hypothesis testing is used to categorize EP similarity. Our hypothesis testing framework begins with the conjecture that aligned cavities with identical binding preferences exhibit a large degree of similarity. Conversely, we also conjecture that aligned cavities with differing binding preferences exhibit a remarkably small degree of EP similarity, when compared to those cavities with identical binding preferences. In accordance with this first premise, our null hypothesis is that the aligned cavities of the pairs of proteins have similar EP fields. Subsequently, our alternate hypothesis is that the aligned cavities of protein pairs have significantly differing EP fields. In other words, the null hypothesis associates variation in EP fields, and their measurements to random chance, while the alternate hypothesis asserts that the variation is significant enough to claim that a non-random cause, i.e. differing binding preferences, causes the disparity in EP fields and their measurements.

The value p, as used in this study, is the probability that an observation would exhibit an EP similarity less than or equal to the EP similarity identified for the specific protein pair (for which the p value is associated) due to chance. In other words, the probability that the observed data would be inconsistent with the null hypothesis, assuming the null hypothesis is true. If the probability is small enough (conventionally .05 is the user defined significance level), then the p value is said to be significant. This observation would justify the rejection of the null hypothesis in favor of the alternate hypothesis that the disparities in EP fields are extreme due to differences in binding preferences.

To carry out this analysis, we first estimate the value of p, which requires us to train a statistical model. We measure EP similarity in every possible combination of proteins with like and non-like binding preferences; this whole set is referred to as E. We then divide the E into training sets and test sets. All training sets are referred to as C, which consist of only like-binding pairs and all test sets (which must be associated to their corresponding training set) are referred to as C' ("not C"), which is a set of all pairs of cavities not in C. C' has both like and non-like binding preference pairs. Additionally, C_i' is an element of the set C'. The model trains on each set C and, if training is successful, it is expected to represent the range of EP similarity measurements that would be anticipated from any other set made up of protein pairs with similar binding preferences.

As it happens, the shape of the frequency distribution of the

EP similarity measurements tightly fits the log-normal distribution. Thus, we use the log normal distribution to estimate the probability, p, of observing a specific value for EP similarity for a given pair of proteins. We make this estimation by approximating the parameters of the log-normal distribution: μ and σ, which are the population mean and population standard deviation for the log-transformed distribution respectively. We approximate these values by calculating the sample mean and standard deviation for every combination of measurements (i.e. for every possible set of C). We, at last, calculate the value of p by using equation 1. The value of p is estimated to be the area under the log-normal curve to the left of C_i', where the area under the curve is equal to 1.

Given the trained statistical model and the estimated p-values, we hypothesize that protein pairs with a relatively high p-value have identical binding preferences, while those with a small p-value have different binding preference. We test this hypothesis in the results section (Section III). When we evaluate our hypothesis, we define true positives (TP) as statistically significant differences between cavity fields that actually have different binding preferences (e.g. between trypsins and chymotrypsins). We define false positives (FN) as statistically significant differences between cavity fields that have the same binding preferences (e.g. between trypsins-trypsins pairs). We define true negatives (TN) as statistically insignificant differences between cavity fields that with the same binding preferences, and false negatives (FN) as statistically significant differences between cavity fields that with the same binding preferences.

C. Building the Data Set

1) Selection: The serine protease superfamily was selected for this study based on the criteria that it exhibits at least two subfamilies with distinct binding preferences. The serine protease class is the best-known class of proteases that uses the classical Ser/His/Asp catalytic triad mechanism, where serine is the nucleophile, histidine is the general base and acid, and the aspartate helps orient the histidine residue and neutralize the charge that develops on the histidine during the transition states. Furthermore, the two subfamilies we choose to study, chymotrypsin and trypsin, are both well-studied members of this class.

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

Fig. 3. **PDB codes used in the data set.**

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 [32] and positively charged amino acids in trypsins [33]. We hypothesize that the electrostatic differences between the S1 subsites of the trypsins and chymotrypsins will be discernible and statistically significant when evaluated with EPAC. Since trypsins prefer positively charged substrates, we evaluated electrostatic distances between all subsites at negative electrostatic thresholds.

2) Preparation: The Protein Data Bank (PDB - 6.21.2011) [34] contains 430 Serine proteases from the chymotrypsin and trypsin families. From this set of proteins, we removed partially disordered structures, mutant structures, and structures with more than 90% sequence identity, with preference for structures associated with publications in print. This filtration resulted in 12 serine protease structures. Within these remaining structures, ions, waters, and other non-protein atoms were removed. Since hydrogens were unavailable in all structures, all hydrogens were removed as well for uniformity. Atypical amino acids (e.g. phosphorylated tyrosines) were not removed. After all hydrogens were removed, all structures were reprotonated using the *reduce* component of MolProbity [35], for uniformity.

3) Alignment: Using Ska [36], an algorithm for aligning protein structures, all serine protease structures were aligned to bovine gamma-chymotrypsin (pdb code: 8gch). All the structures in this superfamily exhibit identical folds, causing the aligned structures of these proteins to exhibit little variation, though smaller variations at the scale of the S1 binding site and other subsites are expected. Following structural alignment, solid representations of binding cavities were generated using a method described earlier [28].

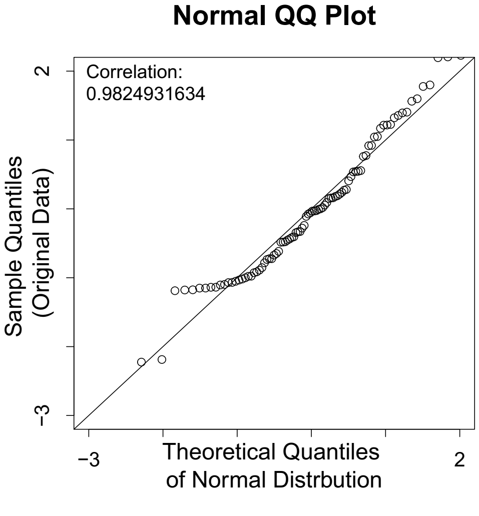

Fig. 4. Quantile-quantile plot of distances between trypsin cavity fields (vertical axis) and the theoretical normal distribution (horizontal axis). This figure evaluates the fit between our data and the log-normal distribution.

III. Experimental Results

A. Evaluating the Statistical Model

We evaluated two parametric models to potentially represent the degree of EP similarity between binding cavities with identical binding preferences. These were the log-normal and

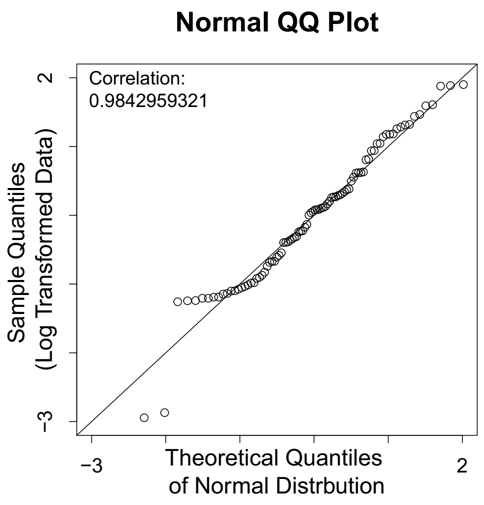

Fig. 5. Quantile-quantile plot of log transformed distances between trypsin cavity fields (vertical axis) and the theoretical normal distribution (horizontal axis). This figure evaluates the fit between our data and the normal distribution.

the normal distributions. To evaluate how well these models fit the variations in electrostatic distances among the trypsin subfamily, we generated quantile-quantile plots.

Figure 4 illustrates the correspondence between the electrostatic distances in our data and the normal distribution. Figure 5 illustrates the correspondence between the electrostatic distances in our data and the log-normal distribution. Note that in Figure 5 we evaluate the log normal distribution by fitting the log-transformed data to the normal distribution. While both distributions fit the data closely, the log-normal distribution exhibited a slightly superior correlation with the data. Visually, it also exhibits a subtly better fit to the data than the normal distribution. Based on these observations, we selected the normal distribution on the log transformed data to estimate p-values.

B. Electrostatic isoPotential Analytical Comparative Model (EPAC)

We used leave-2-out cross-validation to test the predictive accuracy of our model. After we computed the distances between the cavity fields of the serine proteases, we developed a process to designate the training sets and test sets. First, we identified two trypsin proteins that we would leave out of the training set. This process is depicted in Figure 6, where a pair of trypsin are identified from the set of all trypsin-trypsin (t-t) pairs and that pair along with all other pairs with either of these specific trypsins are removed from the training set. Simultaneously, we identified the four pairs of trypsin-chymotrypsin (t-c) that correspond to the chosen t-t pair. After we trained on the condensed training set t-t pairs, we evaluated the statistical significance of the original left out t-t pair and the four selected t-c pairs. CSG differences are asymmetrical, so when the difference of a-b was computed, the difference of b-a was computed as well. This selection and analysis process was repeated until every t-t pair had been left out once.

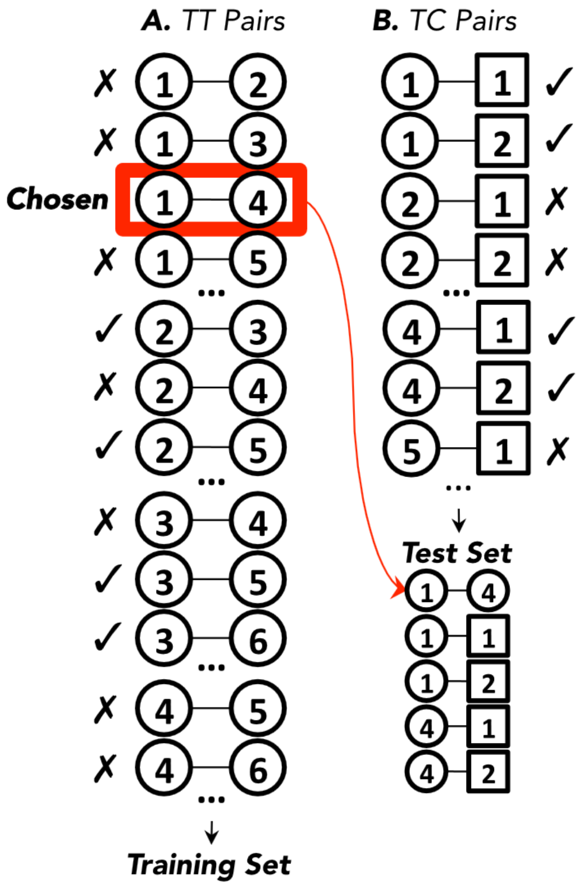

Fig. 6. Circles represent a trypsin and each number within the circle designates a distinct trypsin (1 - 10). Squares represent a chymotrypsin and the numbers 1 and 2 designate the two different chymotrypsin. The line connecting trypsins (circles) and chymotrypsins (squares) indicates pairs and, together, they represent their measurement of EP similarity. This diagram demonstrates how a particular training and testing set would be chosen.

a)	t-c pairs	t-t pairs
Significant	128	12
Insignificant	232	78
b)	t-c pairs	t-t pairs
Significant	360	6
Insignificant	0	84
c)	t-c pairs	t-t pairs
Significant	358	12
Insignificant	2	78
d)	t-c pairs	t-t pairs
Significant	360	18
Insignificant	0	72

Fig. 7. Summary of the total number of TPs, TNs, FPs, and FNs, at the -2.5 kT/e threshold (a), -5.0 kT/e threshold (b), -7.5 kT/e threshold (c), -10.0 kT/e threshold (d).

We evaluated the number of TPs, TNs, FPs, and FNs at four electrostatic potential thresholds: -10.0 kT/e, -7.5 kT/e, -5.0 kT/e, and -2.5 kT/e (Figure 7). Based on the conventional standard of significance, .05, EP measurements from proteins with similar binding preferences (the t-t pairs) were statistically insignificant in the majority of cases. At the smallest threshold we measured at, -10.0 kT/e, the EP similarities for t-t pairs were statistically insignificant in 72 out of 90 cases. For the -7.5 kT/e threshold, t-t pairs were statistically insignificant in 78 out of 90 cases. For the -5.0 kT/e threshold, 84 out of 90 t-t pairs were statistically insignificant and the -2.5 kT/e threshold had 78 out of 90 statistically insignificant t-t pairs. At different potential thresholds, t-t pairs were statistically insignificant at similar frequencies that remained close to 91.4 percent.

In the proteins with different binding preferences (the t-c pairs), EP measurements were statistically significant according to the conventional standard of significance, .05, in the majority of cases, especially when using smaller potential thresholds. At the smallest threshold we measured at, -10.0 kT/e, the EP differences for t-c pairs were statistically significant in 360 out of 360 cases. For the -7.5 kT/e threshold, t-c pairs were statistically significant in 358 out of 360 cases. For the -5.0 kT/e threshold, 360 out of 360 t-c pairs were statistically significant and the -2.5 kT/e threshold had 128 out of 360 statistically significant t-c pairs. Overall, t-c pairs were less frequently statistically significant as electrostatic potential thresholds approached zero. These results indicate that sensitivity improves as potential thresholds reach -10.0 kT/e. This effect is apparent in Figure 8, which illustrates the relative p-values of t-t and t-c pairs.

We observed that the human trypsin 1 (1trn) displayed results inconsistent with the rest of the data we analyzed. After comparing the different trypsins, we were able to identify a unique characteristic of 1trn that we believe is the cause of this irregularity. This specific trypsin has a highly polar amino acid (a phosphorylated tyrosine 151) abutting the binding site that we are studying. This amino acid is affecting the EP field of the 1trn binding site, causing the majority of the errors in the following numbers. For the smallest threshold we measured at, -10.0 kT/e, 0 out of 18 1trn-t pairs were insignificant and 72 out of 72 1trn-c pairs were significant. This indicated that every pair of proteins that included 1trn, at this threshold, had a statistically significant EP similarity measurement. For the -7.5 kT/e threshold, 1trn-t pairs were statistically insignificant in 6 out of 18 cases and 1trn-c pairs were significant in, again, 72 of 72 cases. The -5.0 kT/e threshold had 12 out of 18 statistically insignificant 1trn-t pairs and had 72 out of 72 statistically significant 1trn-c pairs. For the -2.5 kT/e threshold, there were 16 out of 18 statistically insignificant 1trn-t pairs and only 36 of 72 statistically significant 1trn-c pairs. Aside from data generated at the -2.5 kT/e threshold, false positives (t-t pairs that are statistically significant) at other thresholds always involved 1trn. The only false negatives (t-c pairs that are statistically insignificant) below the -2.5 kT/e threshold occurred when 1bzx and 2eek were left out, nudging the difference between 1bzx-1eq9 into insignificance (p = .050516). 1eq9-1bzx had the same p-value. These observations suggest that the electrostatic differences between 1trn and other tyrosines stem from the contribution of the phosphorylated tyrosine to the negative charge in the S1 subsite, which creates the substantial electrostatic differences with other tyrosines. This effect is also apparent in Figure 8.

IV. Discussion

We have described EPAC, a statistical model for evaluating the similarity of electrostatic isopotentials inside ligand binding sites. After considering normal and log-normal distributions, we observed that log-normal distributions better fit the distribution of distances between cavity fields from trypsins. These preliminary results indicate one example that it is possible to model such distributions.

Our experimental results show that EPAC effectively distinguishes pairs of cavity fields that are distant enough to be statistically significant from those that are not. This trend was observed when EPAC identified statistically significant differences between trypsins and chymotrypsins at all electrostatic potential thresholds, though many more were significant at lower isopotential thresholds. Likewise, most cavity fields from pairs of trypsins were statistically insignificant, especially as the absolute value of isopotential thresholds fall. These results indicate that as we adjust isopotential thresholds from -2.5 kT/e to -10.0 kT/e, EPAC became a more sensitive classifier. This behavior is likely a result of the fact that negative isopotential thresholds that are closer to zero necessarily contain isopotentials generated from thresholds that are more negative. Larger isopotentials clearly diminish the specificity of EPAC as a classifier, creating a larger number of false positives (statistically significant differences between trypsin cavity fields). Likewise, as isopotential thresholds fall towards more negative values, false positives dropped. If thresholds were extended even further to the negative side, cavity fields become so small in volume that the probability of intersection even between trypsins begins to fall, creating greater numbers of false negatives. This effect was not observed at the isopotential thresholds we considered.

EPAC also identified differences between the trypsins and the structure of human trypsin 1 (pdb: 1trn). A phosphorylated tyrosine 151 in human trypsin 1 created differences with other trypsin cavity fields that was statistically significant, indicating that even subtle differences such as this posttranslational modification of an amino acid can be identified with EPAC.

EPAC is the first statistical model to operate independently of atomic geometry, enabling a comparative analysis of electrostatic fields without an algorithmic dependence on atomic positions. In addition to operating as an stand-alone classifier, this novel capability enables EPAC to add a nearly orthogonal prediction to larger function annotation and specificity annotation systems that complements existing atom-based statistical models. For example, EPAC can contribute electrostatic information to algorithms for comparing ligand binding sites (e.g. [3]) and for identifying electrostatic influences on specificity (e.g. [30]).

Acknowledgement

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

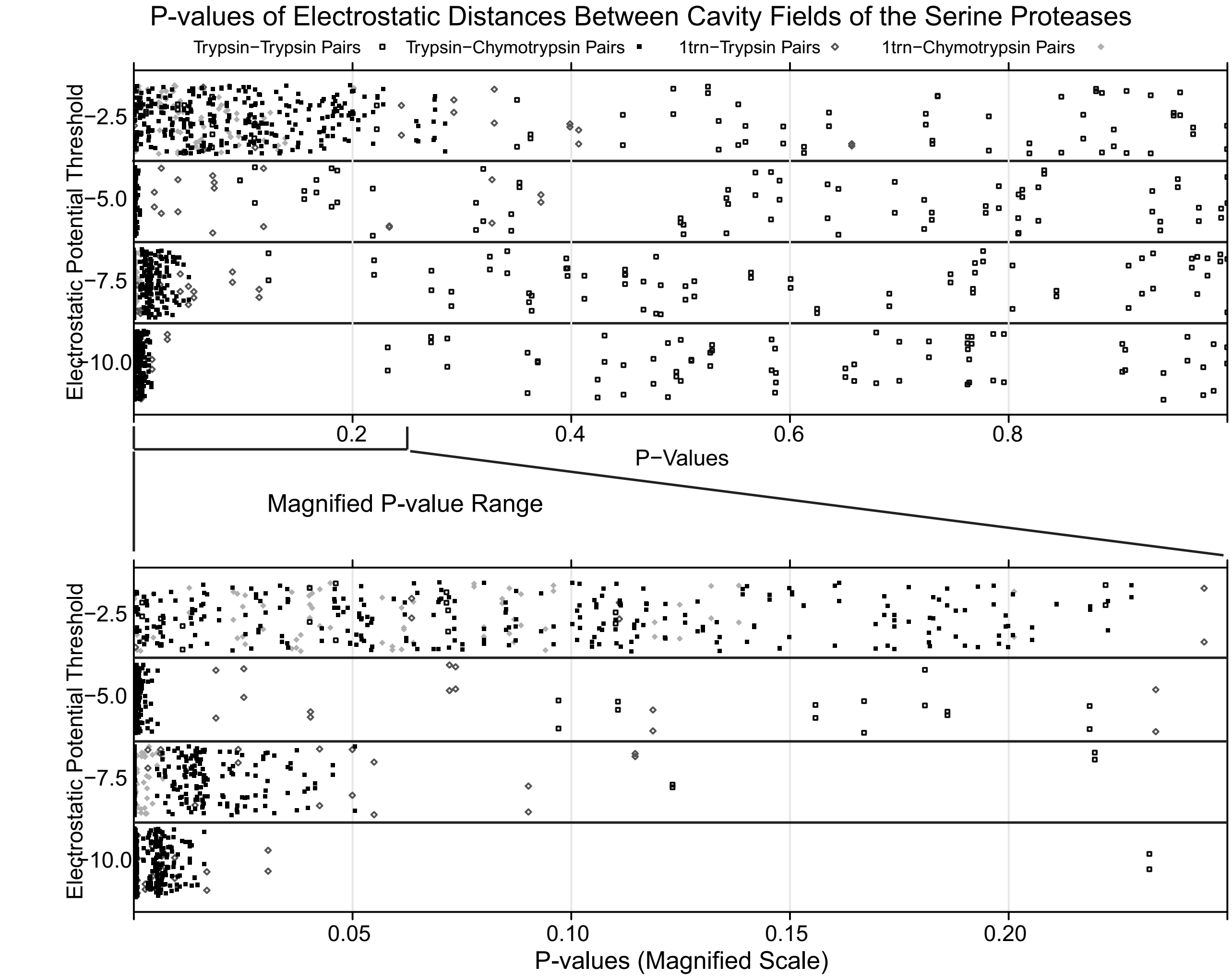

Fig. 8. P-values of electrostatic distances between cavity fields of the serine proteases. The upper graph plots all p-values, whereas the lower graph plots p-values between 0 and .25, for enhanced clarity. In both plots, the horizontal axis plots the p-value of a pair of cavity fields. The vertical axis is divided into four rows that correspond to pairs of cavity fields that have been generated at different isopotential thresholds. Vertical differences between points inside each row are irrelevant and are shown to illustrate density at any given p-value. Open shapes plot the p-value of trypsin-trypsin pairs. Closed shapes plot the p-values of trypsin-chymotrypsin pairs. Grey diamonds plot the p-values of pairs involving 1trn, a trypsin with a phosphorylated tyrosine 151 near the S1 binding site. Black squares plot the p-values of all other pairs.

References

[1] A. Stark, S. Sunyaev, and R. B. Russell, "A model for statistical significance of local similarities in structure," *Journal of molecular biology*, vol. 326, no. 5, pp. 1307–1316, 2003.

[2] T. A. Binkowski, A. Joachimiak, and J. Liang, "Protein surface analysis for function annotation in high-throughput structural genomics pipeline," *Protein Science*, vol. 14, no. 12, pp. 2972–2981, 2005.

[3] B. Chen, "Algorithms for structural comparison and statistical analysis of 3d protein motifs by chen, vy fofanov, dm kristensen, m. kimmel, o. lichtarge, and le kavraki pacific symposium on biocomputing 10: 334-345 (2005)," in *Pacific Symposium on Biocomputing*, vol. 10. Citeseer, 2005, pp. 334–345.

[4] V. Y. Fofanov, B. Y. Chen, D. H. Bryant, M. Moll, O. Lichtarge, L. Kavraki, and M. Kimmel, "A statistical model to correct systematic bias introduced by algorithmic thresholds in protein structural comparison algorithms," in *Bioinformatics and Biomeidcine Workshops, 2008. BIBMW 2008. IEEE International Conference on*. IEEE, 2008, pp. 1–8.

[5] 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*, 2011, pp. 22–9.

[6] ——, "A Statistical Model of Overlapping Volume in Ligand Binding Cavities," in *Proceedings of the Computational Structural Bioinformatics Workshop (CSBW 2011)*, 2011, pp. 424–31.

[7] B. Y. Chen and S. Bandyopadhyay, "Modeling regionalized volumetric differences in protein-ligand binding cavities," *Proteome Sci*, vol. 10, no. Suppl 1, p. S6, 2012.

[8] ——, "A regionalizable statistical model of intersecting regions in protein–ligand binding cavities," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 03, 2012.

[9] 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." *J Comp Biol*, vol. 14, no. 6, pp. 791–816, 2007.

[10] L. Xie and P. E. Bourne, "A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites." *BMC Bioinformatics*, vol. 8 Suppl 4, p. S9, Jan. 2007.

[11] ——, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments," *Proceedings of the National Academy of sciences*, vol. 105, no. 14, pp. 5441–5446, 2008.

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

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

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

[15] P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice, and P. Willett, "A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures," *Journal of molecular biology*, vol. 243, no. 2, pp. 327–344, 1994.

[16] Y. Ye and A. Godzik, "Fatcat: a web server for flexible structure comparison and structure similarity searching," *Nucleic acids research*, vol. 32, no. suppl 2, pp. W582–W585, 2004.

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

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

[19] Y. Y. Tseng, J. Dundas, and J. Liang, "Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns," *Journal of molecular biology*, vol. 387, no. 2, pp. 451–464, 2009.

[20] K. Kinoshita, Y. Murakami, and H. Nakamura, "ef-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W398–W402, 2007.

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

[22] B. Y. Chen, D. H. Bryant, V. Y. Fofanov, D. M. Kristensen, A. E. Cruess, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "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.

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

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

[25] B. Chen, V. Fofanov, B. D.H., B. Dodson, D. Kristensen, A. Lisewski, M. Kimmel, O. Lichtarge, and L. Kavraki, "Geometric sieving: Automated distributed optimization of 3D motifs for protein function prediction," *Proceedings of The Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006)*, pp. 500–15, April 2006.

[26] K. Kinoshita, J. Furui, and H. Nakamura, "Identification of protein functions from a molecular surface database, ef-site," *Journal of structural and functional genomics*, vol. 2, no. 1, pp. 9–22, 2002.

[27] K. Kinoshita and H. Nakamura, "Identification of protein biochemical functions by similarity search using the molecular surface database ef-site," *Protein Science*, vol. 12, no. 8, pp. 1589–1595, 2003.

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

[29] W. Rocchia, E. Alexov, and B. Honig, "Extending the applicability of the nonlinear poisson-boltzmann equation: Multiple dielectric constants and multivalent ions," *The Journal of Physical Chemistry B*, vol. 105, no. 28, pp. 6507–6514, 2001.

[30] B. Y. Chen, "Vasp-e: Specificity annotation with a volumetric analysis of electrostatic isopotentials," *PLoS Comput Biol*, vol. 10, no. 8, 08 2014.

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

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

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

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

[35] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "Molprobity: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 1, pp. 12–21, 2009.

[36] 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.