Journal of Bioinformatics and Computational Biology
Vol. 10, No. 3 (2012) 1242004 (18 pages)

DOI: 10.1142/S0219720012420048

Imperial College Press
www.icpress.co.uk

A REGIONALIZABLE STATISTICAL MODEL OF INTERSECTING REGIONS IN PROTEIN–LIGAND BINDING CAVITIES

BRIAN Y. CHEN[*,‡] and SOUTIR BANDYOPADHYAY[†,§]

[*]*Department of Computer Science and Engineering*
Lehigh University, 19 Memorial Drive West
Bethlehem, PA 18015, USA

[†]*Department of Mathematics, Lehigh University*
14 East Packer Avenue, Bethlehem, PA 18015, USA
[‡]*chen@cse.lehigh.edu*
[§]*sob210@lehigh.edu*

Received 16 February 2012
Revised 7 April 2012
Accepted 7 April 2012
Published 4 June 2012

Finding elements of proteins that influence ligand binding specificity is an essential aspect of research in many fields. To assist in this effort, this paper presents two statistical models, based on the same theoretical foundation, for evaluating structural similarity among binding cavities. The first model specializes in the "unified" comparison of whole cavities, enabling the selection of cavities that are too dissimilar to have similar binding specificity. The second model enables a "regionalized" comparison of cavities within a user-defined region, enabling the selection of cavities that are too dissimilar to bind the same molecular fragments in the given region. We applied these models to analyze the ligand binding cavities of the serine protease and enolase superfamilies. Next, we observed that our unified model correctly separated sets of cavities with identical binding preferences from other sets with varying binding preferences, and that our regionalized model correctly distinguished cavity regions that are too dissimilar to bind similar molecular fragments in the user-defined region. These observations point to applications of statistical modeling that can be used to examine and, more importantly, identify influential structural similarities within binding site structure in order to better detect influences on protein−ligand binding specificity.

Keywords: Protein structure comparison; structural bioinformatics; statistical models; statistical shape analysis.

1. Introduction

Discovering influences on protein−ligand binding specificity is a crucial aspect of research in molecular biology, bioengineering, drug design, and other fields. In such

‡Corresponding author.

settings, painstaking visual examination of protein structures can provide explanations for biochemical observations made in the past, while informing the design of future experiments along basic biophysical principles. As identified by visual examination, regional similarities between binding cavities may bind the same molecular fragment, while regional variations elsewhere may create differences in specificity. Similarities and variations of this nature are potential influences on specificity and, once identified, they point to experiments that examine the extent of their influence.

But visual examination requires expertise in structural biology, and the consideration of many structures is ultimately constrained by human limitations and error. To guide and accelerate these efforts, computational methods can identify potential influences on specificity.[1−3] One approach has been to identify similarities and variations in binding cavity shape. Boolean set operations can be used to detect overlapping and non-overlapping regions in solid representations of binding cavities (Fig. 1). Cavities with large overlapping regions may have similar binding preferences, while cavities with large non-overlapping regions may accommodate different ligands.[1] Such methods can thus assist human efforts because they can automatically separate cavities likely to have similar binding preferences from those likely to be different.[4,5]

This "unified" approach to the comparison of cavities, common among most binding site comparison algorithms (e.g. Refs. 6 and 7), evaluates similarity between entire cavities. But binding cavities can have similarities in some regions and differences in others, causing some regions to have very different impacts on specificity. Unified methods have no means to assess the importance of a user-defined region on specificity. To address this problem, this paper proposes a "regionalized" comparison of protein cavities that detects when two or more cavities are similar enough within a user-defined region to accommodate similar molecular fragments in that region. The position of the regions detected can indicate to a human user that an area contained in detected regions may be responsible for similar specificity.

Boolean set operations offer a unique opportunity to regionalize the comparison of protein−ligand binding cavities because comparative analysis can be focused on any region with Boolean intersections. As we will show in this paper, regionalization enables the construction of statistical models that are trained on the degree of structural variation inside individual regions: Similarity is significant in regions where no pockets are similar, while the same degree of similarity may not be significant in regions where pockets are very similar. Because our regionalized model can be independently trained on different cavity regions, their prediction thresholds are customized to account for differences in structural variability and conservation. In comparison to earlier unified methods, paraphrased here for comparison, regionalized methods add the additional capability of automatically isolating regions within protein cavities that influence specificity.

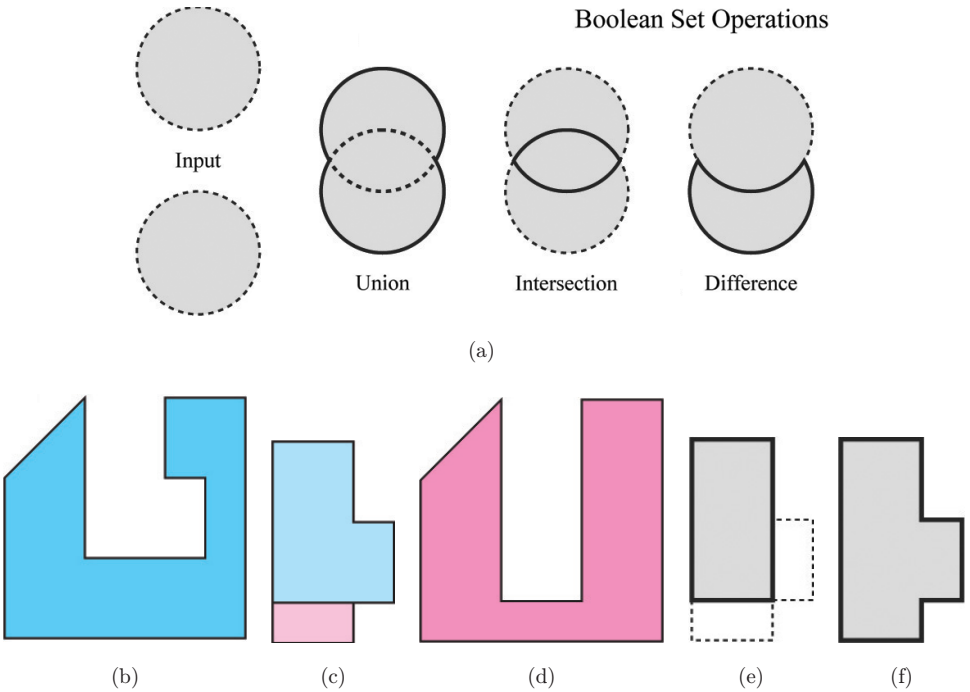

Fig. 1. A diagram of Boolean set operations (a). Aligned proteins with distinctive cavities (b), (d). Overlapping cavities (c). The Boolean intersection (e) and union (f) of the cavities, which is used to compute volumetric similarity.

2. Related Work

The methods described in this paper build on a new approach to protein structure comparison based on volumetric similarities and differences in ligand binding cavities.[1] 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, our earlier work described statistical models for volume-based comparison methods, including the first statistical models of differential volume[4] and overlapping volume.[5]

Point-based representations have notable strengths in comparison efficiency. The least-squares alignment of points in space[8] enables structure comparison software to rapidly consider thousands of atomic superpositions in a database search for the alignment of two or more protein structures with greatest geometric similarity.[9–13] Other approaches to point-based structure alignment, which employ distance matrices[14] and geometric graphs[15,16] are also extremely efficient. These alignment methods inspired the design of newer algorithms for the flexible alignment of protein

structures,[17–19] and fuel the ongoing exploration of the space of protein folds.[20] 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[21] to a more continuous space of variations.[16,22,23]

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,[24–26] sometimes referred to as a *motif*, 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,[26–28] to integrate geometric data from multiple structures,[29–31] and the integration of empty spaces inside binding sites.[32,33] As a result of these developments, motif comparison algorithms can be extremely accurate point-based methods for identifying proteins that catalyze the same reaction.[30]

Surface-based methods use surfaces or surface patches to represent solvent-accessible shapes.[34,35] The surface itself is often described with triangular meshes,[36,37] three-dimensional grids,[38] alpha shapes,[39–41] or spherical harmonics.[42–44] Surface representations have been applied to the comparison of protein structures[36,37] and electrostatic potentials,[45] as well as hybrid representations that combine point-based and surface-based information,[32] but they can also be used to predict the location of binding sites[39,46–48] and hot spots.[49]

Statistical modeling is a critical aspect of both point- and surface-based methods, because it enables an automated and quantitative separation between similar and varying binding sites: Empirical,[40] parametric,[25,50] and nonparametric[51] models can identify pairs of binding 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.[4] In contrast to these existing models, the methods described here model the volume of volumetric overlaps between cavities with identical binding preferences, and extend them to independently model regions inside protein cavities. The result is the first automated method for automatically isolating regions that influence specificity.

3. Methods

In earlier work,[5] we presented a prototype statistical model for identifying statistically significant intersections of protein–ligand binding cavities. Using this model we observed that groups of cavities with different binding preferences exhibit volumetric similarity [Eq. (1)] that is low and statistically significant (i.e. unusual) relative to the higher degree of volumetric similarity found among cavities with similar binding preferences. We first summarize methods related to this model.

We extended our earlier work by showing that we can model volumetric similarity within a user-defined subregion of a set of binding cavities. This regionalized approach enables the statistical significance of an overlapping pair of cavities to be independently evaluated within a specific region, rather than within whole cavities. Depending on the user-defined subregion, the regionalized approach can ignore highly variable regions, while scoring more stringently in conserved regions.

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

3.1. *Computing volumetric similarity*

Given a set of k aligned binding cavities $C = c_1, c_2, \ldots, c_k$, we define the volumetric similarity of these cavities, $d(C)$, using Eq. (1). When evaluating $d(C)$, we first generate intersection ($\cap$) and union ($\cup$) regions with Boolean set operations [Fig. 1(a)] developed in earlier work.[1] We then measure the volume, $v()$, of these regions using the Surveyor's Formula.[52] The geometric interpretation of a set of aligned cavities with high volumetric similarity (e.g. close to 1.0) is that they overlap closely, and thus have very similar shape. Cavities with low volumetric similarity (close to 0.0) overlap poorly.

3.2. *A unified statistical model of volumetric similarity*

Our unified statistical model 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.

We test 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 0.05) then it is hard to reasonably continue assuming that the null hypothesis likely 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 $n > k$ aligned cavities from proteins known to exhibit identical binding preferences. For each of the $\binom{n}{k}$ combinations 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. While the scarcity of protein structure data enabled us to train our models using all combinations, larger training sets can be used without all combinations. These data are represented in a frequency distribution D [see Fig. 3(a)].

It happens that the shape of D tightly fits a *log-normal* distribution, as demonstrated in Sec. 3. The *log-normal* distribution represents an estimate of the distribution, of volumetric distances we might expect if our training data was infinite. Here, we can use it to estimate the probability p of observing a set of k cavities called C', where $d(C')$ is less than that of our input set, $d(C)$, and specificity identical to cavities in T.

We can estimate p by approximating the essential parameters of the *log-normal* distribution: μ and σ, which are the mean and standard deviation for the log-transformed distribution, respectively. We approximate μ and σ with the mean ($\overline{x}$) and standard deviation (s) of the log-transformed sample data, as shown in Eq. (2), where Φ is the cumulative distribution function of the standard normal distribution; p is the proportion of the volume under the log-normal curve to the left of $d(C)$, relative to the total volume under the curve ($x \geq 0$).

$$p(d(C') \leq d(C)) = \Phi\left(\frac{\log d(C) - \mu}{\sigma}\right) \approx \Phi\left(\frac{\log d(C) - \overline{x}}{s}\right). \tag{2}$$

We fit the *log-normal* distribution to D so 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-normal* distribution accurately estimates the underlying probability p, then we can use the *log-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 test this hypothesis in Secs. 4.1 and 4.2.

3.3. *A regionalized statistical model*

The purpose of our regionalized statistical model is to estimate the probability p that two cavities are similar enough within a user-defined region g to bind similar molecule fragments in that region. Any closed region g can be used. To achieve such a model, we first define a regionalized measure of volumetric similarity, $d_g(C)$ for a set of aligned cavities $C = c_1, c_2, \ldots, c_k$. [Eq. (3)].

$$d_g(C) = \frac{v\left(\bigcap_{i=1}^{k}(c_i \cap g)\right)}{v\left(\bigcup_{i=1}^{k}(c_i \cap g)\right)} \tag{3}$$

Three special cases involving regional volumetric similarity can arise. First, $d_g(C)$ can be undefined, because the Boolean union of C inside g may have zero volume. In this case, cavities in C are considered categorically dissimilar. Second, $d_g(C)$ may be zero, in which case the cavities are again categorically dissimilar. Third, $d_g(C)$ may be one, in which case the cavities are considered categorically similar. These special cases are not used for training the model, because they lead to pre-defined conclusions, and if asked to evaluate the p value of a special case, the result is always 0.0 (case 1 or 2) or 1.0 (case 3).

Given our regionalized measure of volumetric similarity, we can build our regionalized statistical model in a manner similar to our unified model: Using $d_g(C)$, we build a regionalized hypothesis testing framework: We assume that regions within aligned cavities that bind similar molecular fragments will exhibit a *large* degree of regional volumetric similarity, and that regions within aligned cavities that bind different molecular fragments will exhibit an *unusually small* degree of regional volumetric similarity.

We test 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 $d_g(C') \leq d_g(C)$. If p is improbably low (typically ≤ 0.05) then we reject the null hypothesis because it seems more probable that $d_g(C)$ is low because the cavities of C bind different molecular fragments in g. To test the null hypothesis, we estimate the probability p based on a training set, T, consisting of $n > k$ aligned cavities from proteins known to bind the same molecular fragments in g. For each of the $\binom{n}{k}$ combinations of members of T, called t, we compute $d_g(t)$, and represent them in a frequency distribution D, which fits tightly with the *log-normal* distribution. We estimate p as the proportion of the volume under the log-normal curve to the left of $d(C)$, relative to the total volume under the curve.

3.4. *Dataset construction and experimental setup*

Protein Families. The serine protease and enolase superfamilies were selected on the criteria that each superfamily exhibit three subfamilies with distinct binding

preferences, and that variations in specificity 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 dataset are derived from the $S1$ subsite, which binds aromatics in chymotrypsins,[53] positively charged amino acids in trypsins,[54] and small hydrophobics in elastases.[55]

Proteins in the enolase superfamily catalyze a reaction that abstracts a proton from carbons adjacent to a carboxylic acid.[56] Opposite an N-terminal "capping domain",[57] 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.[56] Cavities in our dataset, on these amino acids, were classified into three subfamilies that facilitate the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate, in enolase,[58] convert (R)-mandelate to and from (S)-mandelate,[59] in mandelate racemase, and reciprocally cycloisomerize cis,cis-muconate to and from muconolactone, in muconate-lactonizing enzyme.[56] Since members of the Enolase family can exhibit open and closed conformations, only structures with the open conformation were used, for consistency.

Selection. The Protein DataBank (PDB — 6.21.2011)[60] 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 (Fig. 2). Within these structures, ions, waters, and other nonprotein 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.

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 Cycloisomerase: 2pgw, 2zad

Fig. 2. PDB codes of structures used.

Alignment. Using Ska,[13] 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). Both superfamilies exhibit identical folds, leading to nearly perfect alignment of all structures. These alignments are so close that, in earlier work,[1] we observed that alignments to other structures in our dataets generated identical results. Following structural alignment, solid representations of binding cavities were generated using a method described earlier.[1]

Performance. Opteron 6128 processors with 32 GB of random access memory (RAM) were used for all experimentation. Our software is single threaded and requires less than 1GB of RAM. Computing volumetric similarity between a pair of cavities required an average of 0.57 sec. Training and testing on cross-fold validation models required runtimes proportional to the number of combinations considered in each dataset. Leave-2-out tests, for example, required 0:52 sec for enolases, and 1:44 for serine proteases (min:sec), total.

4. Experimental Results

In previous work, we demonstrated that log-normal distributions are an accurate model of the volume of cavity intersections, and that the log-normal model can identify cavities with identical specificity, based on statistically significant volumes of intersection.[5] In this section, we summarize these earlier results and describe related experiments not found in earlier work. We then extend these results to demonstrate that a log-normal distribution can also be applied in a regional context, despite the presence of other categories of data, and finally show that regions with statistically significant overlaps isolate regions of protein cavities with an experimentally established influence on specificity.

4.1. *Validating the log-normal 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 3 illustrates this point on the volumetric similarity between pairs of serine protease cavities as an example: In Fig. 3(b), the log transformed volumetric similarity sample data visibly follows a normal distribution. Furthermore, inspecting the quantile−quantile plots relating the data in Fig. 3(b) to gamma, Weibull, Pareto, Generalized Extreme Value (GEV), and Log-Normal distributions (Figs. 3(c)−3(f)), it is clear that the log-normal and gamma plots are more linear than the others.

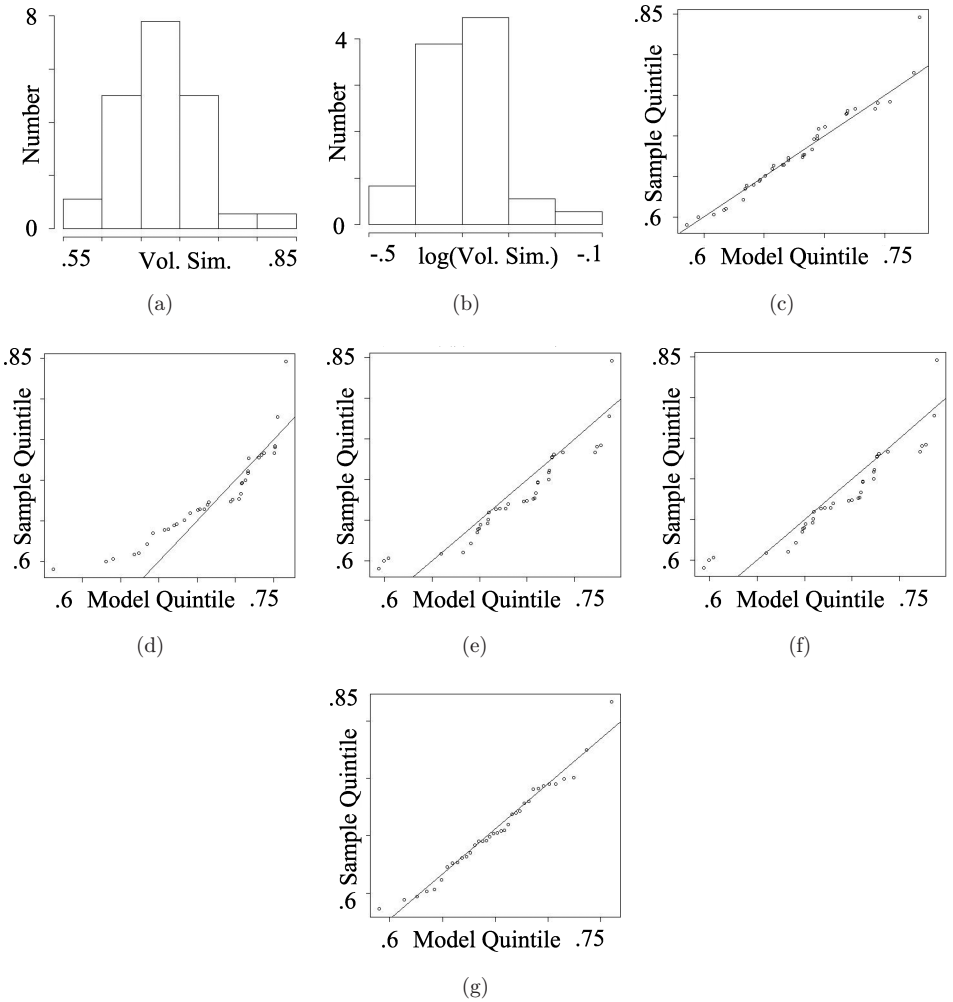

Fig. 3. Volumetric similarity between pairs of trypsin cavities (a), log transformed (b). Quantile-Quantile plots of the gamma (c), weibull (d), Pareto (e), GEV (f), Log-Normal (g) models against the log transformed sample.

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*

Multifold cross-validation was used to fully test the predictive accuracy of our unified model. First, we computed the statistical significance of volumetric similarity among

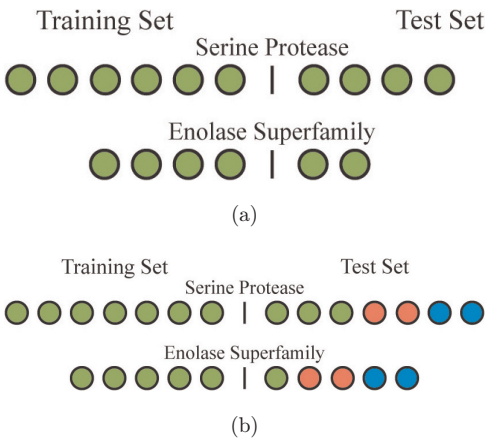

Fig. 4. Multifold cross-validation experimental setup. In serine proteases, circles represent trypsin (green), elastase (red), chymotrypsin (blue) cavities. Among the enolase superfamily, circles represent enolases (green), mandelate racemases (red), and Muconate Cycloisomerase (blue). Cavities were allocated either to the test set or the training set, demonstrating one fold of a leave-4-out (a, top) and one fold of a leave-2-out (a, bottom) comparison of cavities with identical specificities. B demonstrates one fold of a leave-4-out (b, top) and one fold of a leave-2-out (b, bottom) comparison of cavities with differing specificity.

cavities having binding preferences identical to the training set. For both trypsins and enolases, we left out two cavities, while training our unified model on the volumetric similarity of pairs of cavities from the remaining cavities. This is illustrated in the case of enolase, at the bottom of Fig. 4(a). 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, 0.05, 42 out of 45 trypsin validation runs and 13 out of 15 enolase validation runs had statistically insignificant volumetric similarity. Cavities with identical specificity in our dataset were so consistently similar that no pair exhibited volumetric similarity that was statistically significant relative to the others.

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 Fig. 4(a)). 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 generally statistically insignificant.

Next, we examined the ability of our unified model to measure the statistical significance of volumetric similarity among cavities having binding preferences distinct from the training set. For both trypsins and enolases, we left out one cavity, and trained our unified model 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 testing 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 Fig. 4(b). Pairs of cavities with different binding preferences were dissimilar enough that volumetric similarity between them was statistically significant in 91 out of 100 serine protease pairs and 59 out of 60 enolase pairs.

Again, because of the larger size of the trypsin set, we performed leave-2-out and leave-3-out cross validation by training our unified model 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 tested 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 Fig. 4(b). 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.

In general, almost all sets of cavities with identical binding preferences exhibited measures of volumetric similarity that did not differ significantly from all other sets of cavities with identical binding preferences. In contrast, all sets of cavities with differing binding preferences exhibited measures of volumetric similarity that were significantly less than other sets of cavities with identical binding preferences. These results held regardless of the number of cavities in the set considered and for both serine proteases and enolase superfamily cavities.

4.3. *Validating the regional model*

While our regional model is designed to represent the same kind of data as our unified model, it is conceivable that the most appropriate model for this data may not be the log-normal distribution, as observed in Sec. 4.1. As before, we considered the gamma, Weibull, Pareto, GEV, and Log-Normal distributions, and we added the Gaussian and T distributions as possible models for representing the degree of volumetric similarity between binding cavities with identical binding preferences inside a given region g, which was taken to be a cube of sidelength 5.0 Å.

This experiment was repeated for four different g regions from each of the enolase and trypsin training sets, where the number of special cases (as defined in Sec. 3.3) varied from zero to most of set. The GEV and Pareto distributions could not be suitably fit to the data, because estimates of their distribution parameters must be calculated using iterative approaches. These approaches did not stabilize for the two cases where special cases affected a majority of the set, forcing the GEV and Pareto distributions to be eliminated from consideration. In the remaining cases, based again on the comparison of quantile−quantile plots, the log-normal distribution

fit better than the others, except the Gaussian distribution, which performed comparably on this dataset. Given the success of the log-normal distribution on the unified case, and the technical similarity of the unified and regional models, we selected the log-normal distribution for consistency.

4.4. *Testing the regional model*

We tested our regional statistical model by assembling a training set of all enolase and trypsin cavities. From each training set, we excluded one cavity for testing. Among enolases, this was the binding cavity of Enolase 1 from Toxoplasma gondii (3otr), and among serine proteases, this was human trypsin 4 (1h4w). Rather than select an arbitrary user-defined region for analysis, we fully surrounded each training set of aligned cavities with a lattice of cubes having sidelengths of 5.0 Å. Each of the 125 cubes around the enolase training set, and the 252 cubes around the serine proteases was treated individually as a user-defined region for statistical analysis.

Thus, each cube formed the basis for an individually trained regionalized statistical model, as described in Sec. 3.3. The majority of models regionalized in this manner were trivial because the cube where the model was trained intersected with no training set cavities. Most trivial models were created because a generous margin of cubes were generated surrounding the aligned cavities: 65 cubes around the enolase set and 144 cubes around the trypsin set were trivial in this manner.

At every cube, p-values were computed for the regional volumetric similarity between the excluded cavity and the dataset cavities with non-enolase or non-trypsin specificity. Since there were four non-enolases and four nontrypsins in our dataset, four p-values were generated for every cube. Most cubes with nontrivial models exhibited 1 or zero statistically significant p-values, based on our 0.05 significance threshold. These cubes were situated in regions of the cavity alignments where the non-enolase cavities and non-trypsin cavities were essentially identical to enolase and trypsin cavities. That this occurs so frequently is unsurprising because the cavities in both families are strongly defined by the family's overall fold: the TIM-barrel fold in enolases is totally conserved among the entire enolase superfamily, as is the serine protease fold.

Where the cavities do vary, however, many statistically and categorically significant p-values were observed. Seven cubes among the enolase models exhibited four statistically significant p-values, and six cubes among the serine protease models exhibited four statistically significant p-values. A selection of these cubes found on the serine proteases can be seen in Fig. 5. In every case, these cubes corresponded to cavity regions that do not bind the same molecular fragment, as established experimentally by other authors. For example, the four models with the most statistically significant p-values among the serine proteases (Figs. 5A2 B2) correspond to cavity regions essential for accommodating the large hydrophobic sidechains that bind to chymotrypsins.[53]

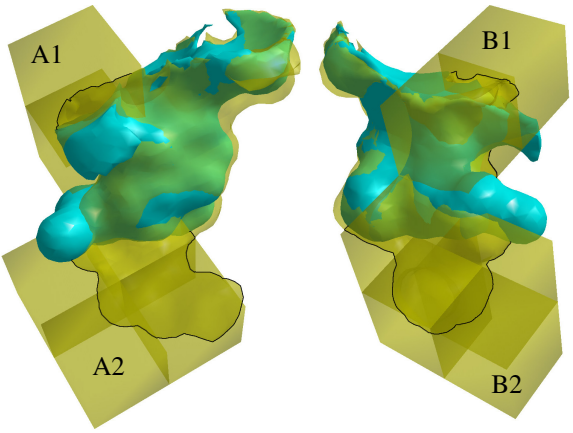

Fig. 5. The structurally aligned S1 cavity of human trypsin 4 (teal) and the S1 cavity of cow chymotrypsin (transparent yellow with black outline), with six regionalization cubes, shown in two orientations. The orientation on the right is that of the left rotated about a near-vertical axis approximately 180°. A1, B1 represent two cubes that generate models with statistically significant p-values. A2,B2 represent four cubes that generate the most statistically significant p-values; these cubes coincide with regions of the chymotrypsin S1 cavity that are essential for accommodating larger and more hydrophobic amino acids that the shorter trypsin cavities cannot accommodate.

5. Conclusions

We have presented a new regionalized statistical model that expanded on a unified statistical model described earlier.[5] To our knowledge, the approach described here is the first method for automatically decomposing multiple structural alignments of protein−ligand binding cavities and evaluating the statistical significance of volumetric similarities within user-defined cavity regions. We demonstrated an application of this regionalized model that is not possible with existing methods: We divided multiple structural alignments of serine protease and enolase cavities into a lattice of cubes and analyzed regionalized volumetric similarity in each cube.

In developing this new statistical model, we observed that the log-normal distribution performed at least as well or better at representing volumetric similarity data in comparison to multiple parametric distributions. In testing the regionalized model, we observed that agglomerations of cubes with statistically significant p-values could identify experimentally established structural influences on ligand binding preferences.

Regionalized statistical models have useful applications where existing models have not been applied, such as in the regionalized analysis of protein−ligand binding cavities for inhibitor design. By identifying regions with a statistically significant lack of similarity among proteins expected to have similar binding preferences, users may be able to identify variations that can be exploited for selective inhibitors. In combination with other sources of biophysical data, regionalized statistical models may thus provide new insights and methods in molecular design and analysis.

Acknowledgments

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

References

1. Chen BY, Honig B, VASP: A volumetric analysis of surface properties yields insights into protein-ligand binding specificity, *PLoS Comput Biol* **6**(8):11, 2010.
2. Nalam MNL, Ali A, Altman MD, Reddy GSKK, Cao H, Gilson MK, Tidor B, Rana TM, Schiffer CA, Evaluating the substrate-envelope hypothesis: Structural analysis of novel HIV-1 protease inhibitors designed to be robust against drug resistance, *J Virol* **84**(10):5368–5378, 2010.
3. Frey K, Georgiev I, Donald B, Anderson A, Predicting resistance mutations using protein design algorithms, *Proc Natl Acad Sci USA* **107**(31):13707–13712, 2010.
4. Chen B, Bandyopadhyay S, VASP-S: A volumetric analysis and statistical model for predicting steric influences on protein-ligand binding specificity, *Proc 2011 IEEE Int Conf Bioinform Biomed (BIBM)*, pp. 22–9, 2011.
5. Chen B, Bandyopadhyay S, A statistical model of overlapping volume in ligand binding cavities, *Proc Comput Struct Bioinform Workshop (CSBW 2011)*, pp. 424–431, 2011.
6. Chen BY, Fofanov VY, Bryant DH, Dodson BD, Kristensen DM, Lisewski AM, Kimmel M, Lichtarge O, Kavraki LE, The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs, *J Comput Biol* **14**(6):791–816, 2007.
7. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, Fofanov VY, Kimmel M, Kavraki LE, Lichtarge O, Prediction of enzyme function based on 3D templates of evolutionarily important amino acids, *BMC Bioinformatics* **9**:17, 2008.
8. Umeyama S, Least-squares estimation of transformation parameters between two point patterns, *IEEE Trans Pattern Analysis and Machine Intelligence* **13**(4):376–380, 1991.
9. Nussinov R, Wolfson HJ, Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques, *Proc Natl Acad Sci USA* **88**(23):10495–10499, 1991.
10. Orengo CA, Taylor WR, SSAP: Sequential structure alignment program for protein structure comparison, *Method Enzymol* **266**:617–635, 1996.
11. Shindyalov IN, Bourne PE, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng* **11**(9):739–747, 1998.
12. Petrey D, Honig B, GRASP2: Visualization, surface properties, and electrostatics of macromolecular structures and sequences, *Method Enzymol* **374**(1991):492–509, 2003.
13. Yang A-S Honig B, An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance, *J Mol Biol* **301**(3):665–678, 2000.
14. Holm L, Sander C, Mapping the protein universe, *Science* **273**(5275):595–603, 1996.
15. Gibrat JF, Madej T, Bryant SH, Surprising similarities in structure comparison, *Curr Opin Struct Biol* **6**(3):377–385, 1996.
16. Xie L, Bourne PE, Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments, *Proc Natl Acad Sci USA* **105**(14):5441–5446, 2008.
17. Ye Y, Godzik A, Flexible structure alignment by chaining aligned fragment pairs allowing twists, *Bioinformatics* **19**(90002) Suppl(2):ii246–i255, 2003.

18. Shatsky M, Nussinov R, Wolfson HJ, FlexProt: Alignment of flexible protein structures without a predefinition of hinge regions, *J Comput Biol* **11**(1):83−106, 2004.
19. Ye Y, Godzik A, Multiple flexible structure alignment using partial order graphs, *Bioinformatics* **21**(10):2362−2369, 2005.
20. Kolodny R, Petrey D, Honig B, Protein structure comparison: Implications for the nature of 'fold space', and structure and function prediction, *Curr Opin Struct Biol* **16**(3):393−398, 2006.
21. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM, CATH−a hierarchic classification of protein domain structures, *Structure* **5**(8):1093−1108, 1997.
22. Krishna SS, Grishin NV, Structural drift: A possible path to protein fold change, *Bioinformatics* **21**(8):1308−1310, 2005.
23. Petrey D, Honig B, Is protein classification necessary? Towards alternative approaches to function annotation, *Curr Opin Struct Biol* **19**(3):363−368, 2009.
24. Russell RB, Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution, *J Mol Biol* **279**(5):1211−1227, 1998.
25. Barker JA, Thornton JM, An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis, *Bioinformatics* **19**(13):1644−1649, 2003.
26. Chen BY, Fofanov VY, Bryant DH, Dodson BD, Kristensen DM, Lisewski AM, Kimmel M, Lichtarge O, Kavraki LE, The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs, *J. Comput. Biol.* **14**(6):791−816, 2007.
27. Polacco BJ, Babbitt PC, Automated discovery of 3D motifs for protein function annotation, *Bioinformatics* **22**(6):723−730, 2006. Available: http://www.ncbi.nlm.nih.gov/pubmed/16410325.
28. Chen BY, Fofanov VY, Bryant DH, Dodson BD, Kristensen DM, Lisewski AM, Kimmel M, Lichtarge O, Kavraki LE, Geometric sieving: Automated distributed optimization of 3D motifs for protein function prediction, *Proc Tenth Annual Int. Conf. Computational Molecular Biology (RECOMB 2006)*, pp. 500−515, 2006.
29. Chen BY, Bryant DH, Cruess AE, Bylund JH, Fofanov VY, Kristensen DM, Kimmel M, Lichtarge O, Kavraki LE, Composite motifs integrating multiple protein structures increase sensitivity for function prediction, *Comput Syst Bioinformatics Conf* **6**:343−355, 2007.
30. Bryant DH, Moll M, Chen BY, Fofanov VY, Kavraki LE, Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction, *BMC Bioinformatics* **11**:242, 2010.
31. Dundas J, Adamian L, Liang J, Structural signatures of enzyme binding pockets from order-independent surface alignment: A study of metalloendopeptidase and nad binding proteins, *J Mol Biol* **406**(5):713−729, 2011. Available: http://www.ncbi.nlm.nih.gov/pubmed/21145898.
32. Chen BY, Bryant DH, Fofanov VY, Kristensen DM, Cruess AE, Kimmel M, Lichtarge O, Kavraki LE, Cavity-aware motifs reduce false positives in protein function prediction, *Proc 2006 IEEE Comput Syst Bioinformatics Conference (CSB 2006)*, pp. 311−323, 2006.
33. Chen BY, Bryant DH, Fofanov VY, Kristensen DM, Cruess AE, Kimmel M, Lichtarge O, Kavraki LE, Cavity scaling: Automated refinement of cavity-aware motifs in protein function prediction, *J Bioinform Comput Biol* **5**(2a):353−382, 2007.
34. Lee B, Richards FM, The interpretation of protein structures: Estimation of static accessibility, *J Mol Biol* **55**(3):379−400, 1971.
35. Connolly M, Solvent-accessible surfaces of proteins and nucleic acids, *Science* **221**(4612):709−713, 1983.

36. Rosen M, Lin SL, Wolfson H, Nussinov R, Molecular shape comparisons in searches for active sites and functional similarity, *Protein Eng* **11**(4):263–277, 1998.
37. Kinoshita K, Nakamura H, Identification of the ligand binding sites on the molecular surface of proteins, *Protein Sci* **14**:711–718, 2005.
38. Laskowski RA, SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions, *J Mol Graph* **13**(5):323–330, 307–308, 1995.
39. Binkowski TA, CASTp: Computed atlas of surface topography of proteins, *Nucleic Acids Res* **31**(13):3352–3355, 2003.
40. Binkowski TA, Adamian L, Liang J, Inferring functional relationships of proteins from local sequence and spatial surface patterns, *J Mol Biol* **332**(2):505–526, 2003.
41. Binkowski TA, Joachimiak A, Protein functional surfaces: Global shape matching and local spatial alignments of ligand binding sites, *BMC Struct Biol* **8**:45, 2008.
42. Ritchie DW, Kemp GJL, Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces, *J Comput Chem* **20**(4):383, 1999.
43. Kazhdan M, Funkhouser T, Rusinkiewicz S, Rotation invariant spherical harmonic representation of 3D shape descriptors, *European Symposium on Geometry Processing 2003*, 2003.
44. Kahraman A, Morris RJ, Laskowski RA, Thornton JM, Shape variation in protein binding pockets and their ligands, *J Mol Biol* **368**(1):283–301, 2007.
45. Zhang X, Bajaj CL, Kwon B, Dolinsky TJ, Nielsen JE, Baker NA, Application of new multi-resolution methods for the comparison of biomolecular electrostatic properties in the absense of global structural similarity, *Multiscale Model Simul* **5**(4):1196–1213, 2006.
46. Nayal M, Honig B, On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites, *Proteins: Struct Funct Genet* **63**:892–906, 2006.
47. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM, A method for localizing ligand binding pockets in protein structures, *Proteins: Struct Funct Bioinf* **62**(2):479–488, 2006.
48. Coleman RG, Sharp KA, Travel depth, a new shape descriptor for macromolecules: Application to ligand binding, *J Mol Biol* **362**(3):441–458, 2006.
49. Bogan AA, Thorn KS, Anatomy of hot spots in protein interfaces, *J Mol Biol* **280**(1):1–9, 1998.
50. Stark A, Sunyaev S, Russell RB, A model for statistical significance of local similarities in structure, *J Mol Biol* **326**:1307–1316, 2003.
51. Chen CP, Posy S, Ben-Shaul A, Shapiro L, Honig B, Specificity of cell-cell adhesion by classical cadherins: Critical role for low-affinity dimerization through beta-strand swapping, *Proc Natl Acad Sci USA* **102**(24):8531–8536, 2005.
52. Schaer J, Stone M, Face traverses and a volume algorithm for polyhedra, *Lecture Notes on Comput Science* **555**(1991):290–297, 1991.
53. Morihara K, Tsuzuki H, Comparison of the specificities of various serine proteinases from microorganisms, *Arch Biochem Biophys* **129**(2):620–634, 1969.
54. Gráf L, Jancsó A, Szilágyi L, Hegyi G, Pintér K, Náray-Szabó G, Hepp J, Medzihradszky K, Rutter WJ, Electrostatic complementarity within the substrate-binding pocket of trypsin, *Proc Natl Acad Sci USA* **85**(14):4961–4965, 1988.
55. Berglund GI, Smalas AO, Outzen H, Willassen NP, Purification and characterization of pancreatic elastase from North Atlantic salmon (Salmo salar), *Mol Mar Biol Biotechnol* **7**(2):105–114, 1998.
56. Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA, The enolase superfamily: A general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids, *Biochemistry* **35**(51):16489–16501, 1996.

57. Rakus JF, Fedorov AA, Fedorov EV, Glasner ME, Hubbard BK, Delli JD, Babbitt PC, Almo SC, Gerlt JA, Evolution of enzymatic activities in the enolase superfamily: L-rhamnonate dehydratase, *Biochemistry* **47**(38):9944−9954, 2008.
58. Kühnel K, Luisi BF, Crystal structure of the Escherichia coli RNA degradosome component enolase, *J Mol Biol* **313**(3):583−592, 2001.
59. Schafer SL, Barrett WC, Kallarakal AT, Mitra B, Kozarich JW, Gerlt JA, Clifton JG, Petsko GA, Kenyon GL, Mechanism of the reaction catalyzed by mandelate racemase: Structure and mechanistic properties of the D270N mutant, *Biochemistry* **35**(18):5662−5669, 1996.
60. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The protein data bank, *Nucleic Acids Res* **28**(1):235−242, 2000.

Brian Y. Chen is currently an Assistant Professor of Computer Science and Engineering at Lehigh University, where his research is focused on algorithms in structural bioinformatics. Before arriving at Lehigh, Dr. Chen was a postdoctoral research scientist at the Howard Hughes Medical Institute, the Department of Biochemistry and Molecular Biophysics, and the Center for Computational Biology and Bioinformatics at Columbia University. He received the Ph.D. in Computer Science from Rice University in 2007 and the M.S. in Computer Science from Rice University in 2003. Dr. Chen received his B.A. in Mathematics and the B.A. in Computer Science from Rutgers University in 2000.

Soutir Bandyopadhyay is currently an Assistant Professor of Mathematics at Lehigh University, where his research is focused on spatial data analysis and time series. He received his Ph.D. in Statistics from Texas A&M University in 2010, his Masters in Statistics from the Indian Statistical Institute in 2005, and his BSc in Statistics from St. Xavier's College, in 2003.