

Journal of Bioinformatics and Computational Biology
© Imperial College Press

A Regionalizable Statistical Model of Intersecting Regions in Protein-Ligand Binding Cavities

BRIAN Y. CHEN*

*Dept. Computer Science and Engineering, Lehigh University,
19 Memorial Drive West
Bethlehem, PA 18015, USA
chen@cse.lehigh.edu*

SOUTIR BANDYOPADHYAY

*Dept. Mathematics, Lehigh University,
14 E. Packer Avenue
Bethlehem, PA 18015, USA
sob210@lehigh.edu*

Received (13 February 2012)

Revised (6 April 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 how influential they are.

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^{1,2,3}. 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 ligands¹. Methods of this sort 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.^{6,7}), 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 bind similar molecule 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 signif-

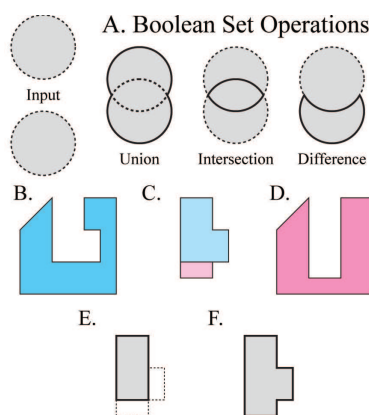


Fig. 1. A diagram of Boolean set operations (A). Aligned proteins with distinctive cavities (B,D). Overlapping cavities (C). The Boolean intersection (E) and union (F) of the cavities, which is used to compute volumetric similarity.

icant 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.

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 cavities¹. 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 volume⁴ and overlapping volume⁵.

Point-based representations have notable strengths in comparison efficiency. The least-squares alignment of points in space⁸ 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^{9,10,11,12,13}. Other approaches to point-based structure alignment, which employ distance matrices¹⁴ 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^{17,18,19}, and fuel the ongoing exploration of the space of protein folds²⁰. As more protein structures become available, the topology of this space, mapped with structure comparison algorithms, appears to be evolving from earlier fold-based clusterings²¹ to a more continuous space of variations^{22,16,23}.

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^{24,25,26}, 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^{27,28,26}, to integrate geometric data from multiple structures^{29,30,31}, 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 reaction³⁰.

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 grids³⁸, alpha shapes^{39,40,41}, or spherical harmonics^{42,43,44}. 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 ^{39,46,47,48} 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 similarities 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 (Equation 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.

3.1. Computing Volumetric Similarity

Given a set of k aligned binding cavities $C = c_1, c_2, \dots, c_k$, we define the volumetric similarity of these cavities, $d(C)$, using Equation 1. When evaluating $d(C)$, we first generate intersection (\cap) and union (\cup) regions with Boolean set operations (Figure 1A) 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.

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

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 .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 Figure 3A).

It happens that the shape of D tightly fits a *log-normal* distribution, as demonstrated in Section 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 (\bar{x}) and standard deviation (s) of the log-transformed sample data, as shown in Equation 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 ($x \geq 0$).

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

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 Sections 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 $C = c_1, c_2, \dots, c_k$. (Equation 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).

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

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 $\leq .05$) 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. Data Set 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 $S4, S3, \dots, S1, S1', S2', \dots, S4'$. 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 chymotrypsins⁵³, positively charged amino acids in trypsin⁵⁴, and small hydrophobics in elastases⁵⁵.

Proteins in the enolase superfamily catalyze a reaction that abstracts a proton from carbons adjacent to a carboxylic acid⁵⁶. 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 reactions⁵⁶. Cavities in our data set, on these amino acids, were classified into three subfamilies that facilitate the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate, in enolase⁵⁸, convert (R)-mandelate to and from (S)-mandelate⁵⁹, in mandelate racemase, and reciprocally cycloisomerize cis,cis-muconate to and from muconolactone, in muconate-lactonizing enzyme⁵⁶. 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

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.

in 14 serine protease and 10 enolase structures (Figure 2). Within these structures, ions, waters, and other non-protein atoms were removed. Since hydrogens were unavailable in all structures, all hydrogens were removed for uniformity. Atypical amino acids (e.g. selenomethionines) were not removed.

Alignment. Using Ska¹³, 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 work¹, we observed that alignments to other structures in our data sets generated identical results. Following structural alignment, solid representations of binding cavities were generated using a method described earlier¹.

Performance. Opteron 6128 processors with 32GB 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 .57 seconds. 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 seconds 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 intersection⁵. 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

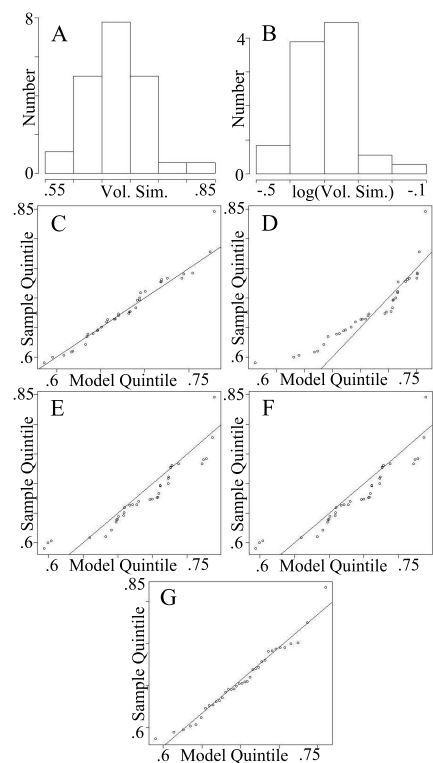


Fig. 3. Volumetric similarity between pairs of trypsin cavities (A), log transformed (B). Quantile-Quantile plots of the gamma (C), weibull (D), Pareto (E), GEV (F), Log-Normal (G) models against the log transformed sample.

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 Figure 3B, the log transformed volumetric similarity sample data visibly follows a normal distribution. Furthermore, inspecting the quantile-quantile plots relating the data in Figure 3B to gamma, Weibull, Pareto, Generalized Extreme Value (GEV), and Log-Normal distributions (Figures 3C-F), it is clear that the log-normal and gamma plots are more linear than the others.

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

4.2. Classifying Cavity Similarity

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 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 Figure 4A. We then evaluated the statistical significance of the volumetric similarity of the left out pair. This process was repeated until every pair of cavities had been left out once. Based on the conventional standard of significance, .05, 42 out of 45 trypsin validation runs and 13 out of 15 enolase validation runs had statistically insignificant volumetric similarity. Cavities with identical speci-

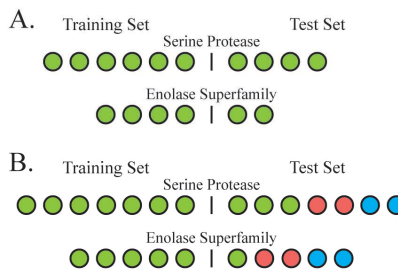


Fig. 4. Multifold cross-validation experimental setup. In serine proteases, circles represent trypsin (green), elastase (red), chymotrypsin (blue) cavities. Among the enolase superfamily, circles represent enolases (green), mandelate racemases (red), and Muconate Cycloisomerase (blue). Cavities were 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.

ficity 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 Figure 4A). 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 Figure 4B. 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 Figure 4B. 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 Section 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 Section 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 Section 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 non-trypsins in our dataset, four p -values were generated for every cube. Most cubes with non-trivial models exhibited 1 or zero statistically significant p -values, based on our .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 Figure 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 (Figure 5A2,B2) correspond to cavity regions essential for accommodating the large hydrophobic sidechains that bind to chymotrypsins⁵³.

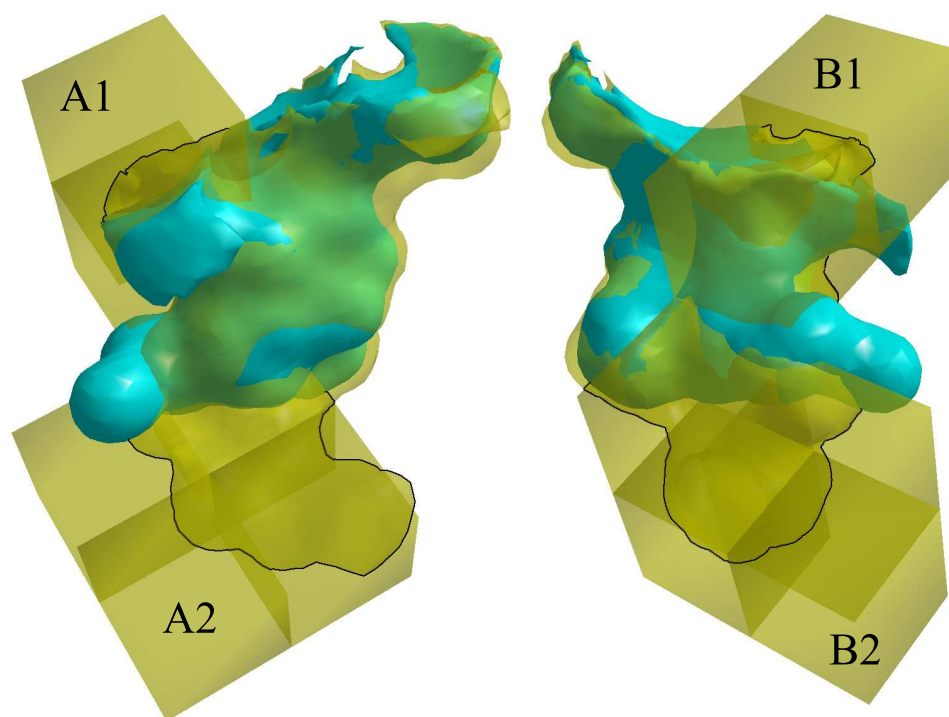


Fig. 5. The structurally aligned S1 cavity of human trypsin 4 (teal) and the S1 cavity of cow chymotrypsin (transparent yellow with black outline), with six regionalization cubes, shown in two orientations. The orientation on the right is that of the left rotated about a near-vertical axis approximately 180 degrees. 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 earlier⁵. 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 start up funds from Lehigh University.

References

1. 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.
2. 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.
3. K. Frey, I. Georgiev, B. Donald, and A. Anderson, “Predicting resistance mutations using protein design algorithms,” *Proc Natl Acad Sci U S A*, vol. 107, no. 31, pp. 13 707–12, 2010.
4. B. Chen and S. Bandyopadhyay, “VASP-S: A Volumetric Analysis and Statistical Model for Predicting Steric Influences on Protein-Ligand Binding Specificity,” in *Proceedings of 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2011, pp. 22–9.
5. —, “A Statistical Model of Overlapping Volume in Ligand Binding Cavities,” in

14 *Brian Y. Chen and Soutir Bandyopadhyay*

- Proceedings of the Computational Structural Bioinformatics Workshop (CSBW 2011)*, 2011, pp. 424–31.
6. 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 Comput Biol*, vol. 14, no. 6, pp. 791–816, 2007.
 7. D. M. Kristensen, R. M. Ward, A. M. Lisewski, S. Erdin, B. Y. Chen, V. Y. Fofanov, M. Kimmel, L. E. Kavraki, and O. Lichtarge, “Prediction of enzyme function based on 3D templates of evolutionarily important amino acids.” *BMC Bioinformatics*, vol. 9, p. 17, Jan. 2008.
 8. S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.
 9. 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.
 10. C. A. Orengo and W. R. Taylor, “SSAP: Sequential Structure Alignment Program for Protein Structure Comparison,” *Method Enzymol*, vol. 266, pp. 617–635, 1996.
 11. 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.
 12. 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.
 13. 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.
 14. L. Holm and C. Sander, “Mapping the protein universe.” *Science*, vol. 273, no. 5275, pp. 595–603, Aug. 1996.
 15. 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.
 16. 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.
 17. Y. Ye and A. Godzik, “Flexible structure alignment by chaining aligned fragment pairs allowing twists.” *Bioinformatics*, vol. 19 Suppl 2, no. 90002, pp. ii246–i255, 2003.
 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. Y. Ye and A. Godzik, “Multiple flexible structure alignment using partial order graphs,” *Bioinformatics*, vol. 21, no. 10, pp. 2362–2369, 2005.
 20. R. Kolodny, D. Petrey, and B. Honig, “Protein structure comparison: implications for the nature of ‘fold space’, and structure and function prediction.” *Curr Opin Struct Biol*, vol. 16, no. 3, pp. 393–8, Jun. 2006.
 21. C. a. Orengo, a. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, “CATH—a hierarchic classification of protein domain structures.” *Structure*, vol. 5, no. 8, pp. 1093–108, Aug. 1997.
 22. S. S. Krishna and N. V. Grishin, “Structural drift: a possible path to protein fold change.” *Bioinformatics*, vol. 21, no. 8, pp. 1308–10, Apr. 2005.

23. D. Petrey and B. Honig, "Is protein classification necessary? Towards alternative approaches to function annotation," *Curr Opin Struct Biol*, vol. 19, no. 3, pp. 363–368, 2009.
24. 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.
25. 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.
26. 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.
27. B. J. Polacco and P. C. Babbitt, "Automated discovery of 3d motifs for protein function annotation." *Bioinformatics*, vol. 22, no. 6, pp. 723–730, 2006. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/16410325>
28. B. Y. Chen, V. Y. Fofanov, D. H. Bryant, B. D. Dodson, D. M. Kristensen, A. M. Lisewski, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Geometric Sieving : Automated Distributed Optimization of 3D Motifs for Protein Function Prediction," in *Proceedings of The Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006)*, 2006, pp. 500–515.
29. B. Y. Chen, D. H. Bryant, A. E. Cruess, J. H. Bylund, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Composite motifs integrating multiple protein structures increase sensitivity for function prediction." *Comput Syst Bioinformatics Conf*, vol. 6, pp. 343–55, Jan. 2007.
30. D. H. Bryant, M. Moll, B. Y. Chen, V. Y. Fofanov, and L. E. Kavraki, "Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction," *BMC Bioinformatics*, vol. 11, no. 242, 2010.
31. J. Dundas, L. Adamian, and J. Liang, "Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and nad binding proteins." *Journal of Molecular Biology*, vol. 406, no. 5, pp. 713–729, 2011. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/21145898>
32. 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.
33. —, "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.
34. 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.
35. M. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *Science*, vol. 221, no. 4612, pp. 709–713, Aug. 1983.
36. 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.
37. 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.
38. 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.
39. T. A. Binkowski, "CASTp: Computed Atlas of Surface Topography of proteins," *Nu-*

- cleic Acids Res*, vol. 31, no. 13, pp. 3352–3355, Jul. 2003.
40. 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.
 41. 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.
 42. 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.
 43. M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, “Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors,” in *European Symposium on Geometry Processing 2003*, 2003.
 44. 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.
 45. 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 absence of global structural similarity,” *Multiscale Model Simul*, vol. 5, no. 4, pp. 1196–1213, 2006.
 46. 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.
 47. 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.
 48. 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.
 49. 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.
 50. 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.
 51. C. P. Chen, S. Posy, A. Ben-Shaul, L. Shapiro, and B. Honig, “Specificity of cell-cell adhesion by classical cadherins: Critical role for low-affinity dimerization through beta-strand swapping.” *Proc Natl Acad Sci U S A*, vol. 102, no. 24, pp. 8531–6, Jun. 2005.
 52. J. Schaer and M. Stone, “Face traverses and a volume algorithm for polyhedra,” *Lect Notes Comput Sc*, vol. 555/1991, pp. 290–297, 1991.
 53. 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.
 54. 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.
 55. 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.
 56. 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.
57. 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.
 58. 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.
 59. 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.
 60. 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.



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 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 Doctor of Philosophy in computer science from Rice University in 2007 and the Master of Science in computer science from Rice University in 2003. Dr. Chen received the Bachelor of Arts in Mathematics and the Bachelor of Arts 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 the Doctor of Philosophy in statistics from Texas A&M University in 2010, the Master of Statistics from the Indian Statistical Institute in 2005, and a Bachelor of Science in Statistics from St. Xavier's College, in 2003.