Isolating Influential Regions of Electrostatic Focusing in Protein and DNA Structure

Seth Blumenthal, Yisheng Tang, Wenjie Yang, and Brian Chen, *Member, IEEE*

Abstract—Electrostatic focusing is a general phenomenon that occurs in cavities and grooves on the molecular surface of biomolecules. Narrow surface features can partially shield charged atoms from the high dielectric solvent, enhancing electrostatic potentials inside the cavity and projecting electric field lines outwards into the solvent. This effect has been observed in many instances and is widely considered in the human examination of molecular structure, but it is rarely integrated into the digital representations used in protein structure comparison software.

To create a computational representation of electrostatic focusing that is compatible with structure comparison algorithms, this paper presents an approach that generates three dimensional solids that approximate regions where focusing occurs. We verify the accuracy of this representation against instances of focusing in proteins and DNA. Noting that this representation also identifies thin focusing regions on the molecular surface that are unlikely to affect binding, we describe a second algorithm that conservatively isolates larger focusing regions. The resulting three dimensional solids can be compared with Boolean Set operations, permitting a new range of analyses on the regions where electrostatic focusing occurs. They also represent a novel integration of molecular shape and electrostatic focusing into the same structure comparison framework.

Index Terms—Electrostatic Focusing, Protein Structure Comparison, Constructive Solid Geometry, Structural Bioinformatics

1 Introduction

Electric fields have a significant influence on molecular recognition [42], [59], [34], [22], [45], [18], [35], [50], [21], [36]. At long atomic distances, their attractive and repulsive forces can affect affinity between interacting molecules [19], [63], [58], [57], [60]. The position of charged atoms can also govern specificity [19], [53], [38] by destabilizing interactions with potential binding partners that lack complementing charges in opposite positions [25], [32]. For the analysis or comparison of static protein structures, an accurate representation of electrostatic potential is thus essential for detecting influences on affinity and specificity by quantitative means [27].

While structure comparison algorithms often incorporate characterizations of electrostatic potentials, *electrostatic focusing* [21], an important nuance of the field, is often omitted. Electrostatic focusing occurs when narrow cavities on the molecular surface partially shield charged atoms from the high dielectric solvent, enhancing potentials inside the cavity and focusing the electric field lines outward. Charged residues along the interior of a cavity can thus have enhanced interactions with other molecules in the cavity than they might have elsewhere. This effect is general and it can be crucial for binding and specificity: It has been observed in proteins, like superoxide dismutase [21] and trypsin [40], as well as in the minor groove of DNA [46]. To enable regions of electrostatic focusing to be explicitly examined, this paper integrates the analysis of electrostatic focusing into the comparison of protein structure and the discovery of binding sites.

In earlier work, we demonstrated that three dimensional solids could be used to represent and compare regions of electrostatic focusing [11]. Applying this approach to the analysis of protein and DNA structures, we verified, using Boolean set operations [10] (Figure 1), that it correctly identified focusing regions, in proteins and DNA, discussed in the literature. However, our method also identified many small regions scattered over the molecular surface where focusing is less influential. To isolate influential regions alone, this paper extends the earlier method by describing a volumetric algorithm and statistical model that separates large focusing regions from false positives that are too small to influence other molecules.

Our results first describe the accuracy of our methods on case studies of superoxide dismutase (SOD) and trypsin, and at a large scale on 866 protein-DNA complexes. We then show how the detection of large focusing regions can point to the locations of ligand binding sites on proteins. On DNA, selecting large focusing regions can isolate locations in the minor groove that interact with amino acid sidechains. Together, our results demonstrate how electrostatic focusing and solid representations can be used together to enable new analytical capabilities. These capabilities have wide applications in protein

- *B. Chen is with the Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, 18015. E-mail: chen@cse.lehigh.edu*
- *S. Blumenthal, Y. Tang, and W. Yang are with Lehigh University.*

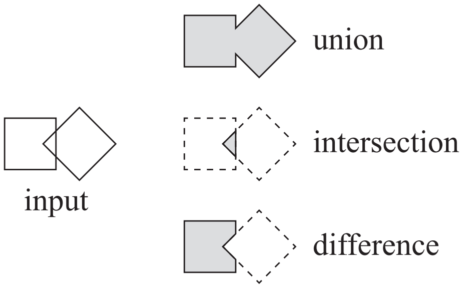


Fig. 1. Boolean Set operations on three dimensional solids (left) can identify regions that are occupied by several solids (union), regions conserved by two solids (intersection), and regions inside one solid and not inside the other (difference). Methods in this paper use Boolean intersections to verify that our methods identify focusing regions with intersecting with intersecting amino acids. .

engineering, drug design, and for determining specific influences on molecular recognition.

2 Related Work

Electrostatic potential has been incorporated widely into protein structure comparison algorithms. The most common approach to this kind of integration has been to add electrostatic labels to existing geometric entities, such as points in space [37], [52], [39], [8], [48], [51], [49], which could represent atoms or amino acids. Similar labelling approaches are equally possible with distance matrices [20] or geometric graphs [16], [62], because label assignment is made to a discrete entity. All these methods correspondences between atoms or amino acids with similar electrostatic qualities. Once a set of corresponding entities has been discovered, least-squares methods [23], [56] are used to measure geometric similarity. In these applications, labeling does not directly effect the final measurement, which is strictly geometric, but rather how the measurement is made (e.g. [9]).

The solvent-accessible surface [31], [13], often referred to as the *molecular surface*, is also frequently used as a digital representation in protein structure comparison algorithms. The molecular surface has been described using polyhedral meshes [47], [28], three dimensional lattices [30], alpha shapes [4], [14], and spherical harmonics [43], [26], [24]. Electrostatic potentials can be mapped onto these surfaces [28], generating a cross-section of the field that is much more precise than atomic or amino acid labels. Surface points are especially crucial because they represent the field at interfaces that make contact with other molecules.

By representing electrostatic information in atom identities or charges, or as potentials on the molecular surface, existing algorithms overlook any description of the field outside the protein, where it influences other molecules. One component of this omission is the effect of electrostatic focusing, where protein shape can enhance potentials in narrow cavity regions outside the protein. For example, in superoxide dismutase, without accounting for electrostatic focusing, there is no "window" of electrostatic potential by which charged substrates can be attracted to the active site from long distances [29]. Existing methods cannot anyway represent this window, but after the enhancement caused by electrostatic focusing is accounted for, the presence of this window is obvious. The isopotential window can be easily visualized with software like GRASP [39], but focusing regions have not been incorporated into algorithms for structure comparison.

Here, we describe an initial approach for incorporating electrostatic focusing into the comparative analysis of molecular structure and the automated identification of binding sites. We develop solid representations that delineate the regions where potentials are enhanced, permitting the geometry of the focusing region to be compared with Boolean set operations. Unlike existing methods, our representation is entirely independent of labeling techniques, enabling focusing regions to be represented outside the molecular surface.

3 Methods

In earlier work [11], we described a method for representing the region of electrostatic focusing. To evaluate this method on a large scale using DNA structures, we also developed a technique for finding amino acids in contact with the minor groove in DNA, and evaluating their position relative to the focusing region. For completeness, we paraphrase these methods here, and describe a new algorithm for separating large focusing regions from thinner regions that are unlikely to have a large influence on binding.

3.1 A Solid Representation of Focusing Regions

Our approach applies Marching Cubes [33] to generate a closed triangular mesh that defines the boundary of the focusing region. As input, it begins with a structure file from the Protein Data Bank (PDB) [3], a threshold K that defines the degree of focusing (used throughout this paper), and a resolution parameter (.5 Å in this work) that controls the geometric detail of the output.

First, we evaluate the field of electrostatic potential surrounding the input molecule with uniform and nonuniform dielectric models. The field computed with a uniform dielectric, Φ_U, models the interior of the input molecule with the same high dielectric as the surrounding solvent. As a result, potentials in Φ_U are not influenced by the shape of the input molecule: no focusing occurs in Φ_U. In contrast, the field computed with a nonuniform dielectric, Φ_N,

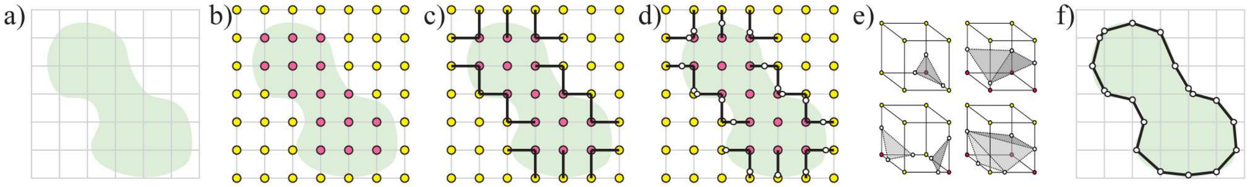


Fig. 2. Representing the Focusing Region. (a) The lattice (grey), surrounding the focusing region (teal). (b) Interior points (red), exterior points (yellow). (c) Edges that connect an interior to an exterior point (bold) (d) boundary points (white). (e) Partial look up table of triangle configurations. (f) Output triangles (dark lines) that approximate the boundary of the focusing region.

represents the interior of the input molecule as having a low dielectric, causing focusing to occur. At any a given point in space p, we can evaluate $f(p)$, the degree of electrostatic focusing at p, with the expression $f(p) = |\Phi_U(p) - \Phi_N(p)|$.

We use DelPhi [44] to compute the two potential fields corresponding to uniform and nonuniform potential fields mentioned above. For the nonuniform dielectric model, we assign an intramolecular dielectric constant of 4 [54], and for the uniform model, we assign 79.9. We use a $0.145\,M$ salt concentration and the AMBER [41] force field for partial charges and atomic radii. Three focusing steps are performed with Delphi on all structures.

Once we have computed both potential fields, we apply Marching Cubes to construct a triangular mesh that approximates the three dimensional boundary of the focusing region. First, we create a three dimensional axis aligned cubic lattice that encloses the entire molecule (Fig. 2a). The lattice can be described as a collection of *lattice points*, *lattice segments* and *lattice cubes*: Lattice points are placed incrementally along each axis, forming a grid. The incremental spacing is set by output resolution. Adjacent lattice points on the same axis define a lattice segment. 12 adjacent lattice segments form a lattice cube.

Second, we determine if each lattice point is inside the region of focusing. A point p is inside the region of focusing if $f(p) > K$, otherwise it is outside, as illustrated in Fig. 2b.

Third, we find every lattice segment that has one point inside and another point outside the region of focusing (Fig. 2c). On each of these segments, we find the point where $f(p) = K$, using linear interpolation. These “boundary points” approximate the limit of the focusing region (Fig. 2d) along each segment.

In the final step, we approximate the surface bounding the region of focusing. Note that for every lattice cube, it’s eight corners must be either inside or outside the focusing region, creating up to 256 inside-outside permutations. We use a look up table (e.g. Fig. 2e), detailed elsewhere [33], to connect each permutation to a triangular approximation of the boundary surface as it passes through the cube. Positioning the corners of each triangle at the boundary points computed above, we store every triangle associated with each lattice cube. The collection of all triangles from all cubes forms a surface that approximates the region of focusing (Fig. 2f).

3.2 Representing the Minor Groove

Here we describe a method for analyzing protein-DNA complexes to identify amino acids that are positioned inside the minor groove. Our approach computes a curving ribbon that connects the phosphate atoms over the minor groove. We use this ribbon to define the border between the interior and the exterior of the minor groove. Amino acids that intersect this ribbon are considered inside the minor groove.

Our method begins with a PDB structure containing a protein-DNA complex. Separating the DNA (Fig. 3a) from the protein, we use the find-pair program from X3DNA [64] to identify nucleotide pairs in the DNA. Unpaired nucleotides, which cannot contribute to a minor groove, are ignored. Separating the input DNA into antiparallel chains, we find the phosphate atom in each nucleotide (Fig. 3b). The center of each phosphatd atom is used as a *control point* for a Catmull-Rom spline [5].

Catmull-Rom splines are a class of piece-wise functions that specify the position of a point along a curve using a number of control points. For any Catmull-Rom spline $c(t)$,

$$c(t) = [1\ t\ t^2\ t^3] \begin{bmatrix} 0 & 2 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 2 & -5 & 4 & -1 \\ -1 & 3 & -3 & 1 \end{bmatrix} \begin{bmatrix} P_{i-2} \\ P_{i-1} \\ P_i \\ P_{i+1} \end{bmatrix}$$

where the spline is parametrized by the variable t and interpolates between control points P_i. By this equation, the segments at the end of each spline cannot be calculated, so we linearly extrapolate a “virtual phosphate” beyond the end of the phosphate backbone in each direction. The end result is a curve without sharp corners that fluidly traces through every phosphate in the phosphate backbone (Fig. 3c).

Once the splines are defined, we begin generating triangles to define the surface of the ribbon. Between consecutive control points, we place five intermediate

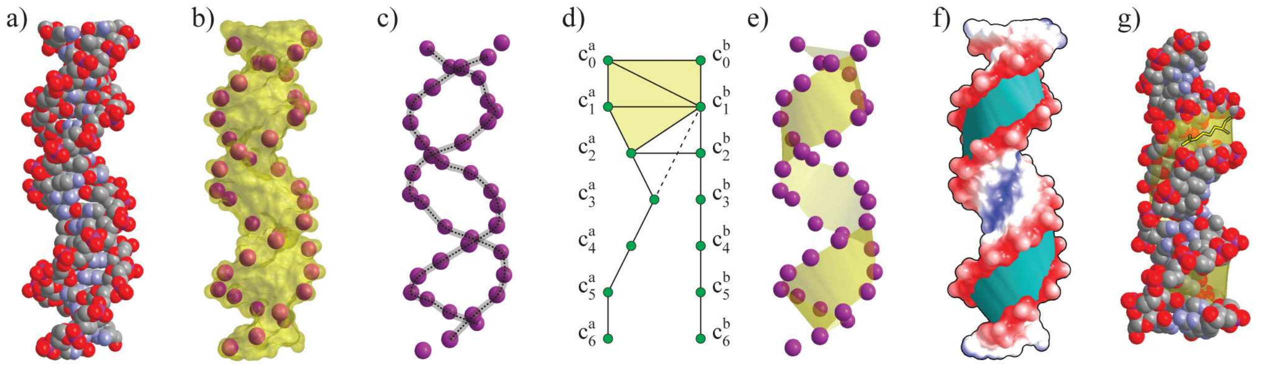


Fig. 3. Finding amino acids in the minor groove. (a) Atoms from DNA chains in 1HLZ, from the PDB (spheres). (b) Molecular surface of DNA (transparent yellow), phosphates in DNA backbone (magenta). (c) Catmull-Rom spline (grey). (d) Diagram of triangle ribbon assembly, corner points (green circles), triangle edges (black lines). (e) Finished ribbon (transparent yellow) (f) Finished ribbon (teal). On the molecular surface of DNA (black outline), blue regions are distant, white regions intermediate, red regions are nearby the ribbon. (g) Atomic structure of DNA, rotated (spheres). Ribbon (transparent yellow), arginine intersecting the ribbon and the minor groove (yellow, black outline).

points along the spline. If we began with k control points along a spline a, combining the control points with the intermediate points results in $6k$ - 5 *corner points*, which we will refer to as $c_0^a, c_1^a, ...c_{6k-6}^a$. This procedure results in two splines, one for each of the phosphate backbones in the complex, and their associated corner points. We refer to these splines as a and b. If the corner points of a and b are enumerated in an antiparallel fashion, we reverse the indices of b to create a parallel enumeration.

Next, we begin generating triangles between the splines: Beginning with the first corner points of each spline, c_0^a and c_0^b, on one end of the DNA. We mark c_0^a and c_0^b *used*. We then measure the distance d_a between points c_1^a and c_0^b and the distance d_b between c_0^a and c_1^b. If d_a is smaller than d_b, we designate c_1^a the *next point*, otherwise this designation is assigned to c_1^b. For example, in Fig. 3d, the distance between c_3^a and c_1^b is larger than the distance between c_2^a and c_2^b, causing us to select c_2^b as the next point. The next point defines a new triangle using the last two used points and the next point. We add the new triangle to a list of triangles that defines the ribbon, and then we also mark the next point used. This triangle generation process then repeats with the last two used points.

The final result is a ribbon of edge-adjacent triangles (see Fig. 3e) that minimize the length of the edges that extend between the splines. This minimization causes the triangles to always cover the minor groove (e.g. Fig. 3f), since the minor groove is typically narrower than the major groove. In theory, however, if the DNA is in a highly nonstandard conformation, the ribbon could switch to the opposite groove. That effect was not observed in our data.

We define an amino acid as being inside the minor groove if it has at least one atom that intersects the ribbon. We detect atom-ribbon intersections by measuring the distance from the center of the atom to any triangle of the ribbon. If the distance is smaller than Van der Waals radius of the atom (based on [6]), we say that the ribbon and the atom intersect (e.g. the arginine on Fig. 3g).

3.3 Isolating Large Focusing Regions

Boundary effects, between the low dielectric interior of a protein and the high dielectric solvent, enhance electrostatic potentials near the molecular surface. Enhancement typically occurs at thin, ubiquitous patches on the molecular surface, as illustrated in Figures 6 and 9, but in narrow cavities, focusing regions can be much larger and more influential. As a collection of patches that nearly cover the molecular surface, the focusing regions we identify in Section 3.1 are nonspecific markers of potential binding sites. To eliminate them, we describe a novel technique for distinguishing large focusing regions from many thin patches, using Boolean set operations (Figure 4).

We begin with two regions defined by polyhedral boundary surfaces: The focusing surface F, as generated in Section 3.1, and the molecular surface, M. At each point p on F, we identify the shortest distance to any triangle of M, and vice versa. If the distance is greater than 2 Å, we place a sphere, centered on p, of radius 2 Å. The result is a large collection of spheres, centered on some of the points of F, that are distant from the molecular surface. This process is repeated on M, generating spheres on points of M that are further than 2 Å from F. Given that the 92 percent of triangles have a sidelength of less than .32 Å, and that the maximum sidelength is approximately .8 Å, the spheres overlap densely in regions where the focusing surface diverges from the molecular surface. In thin regions where the focusing surface and the molecular surface are close by, spheres do not aggregate.

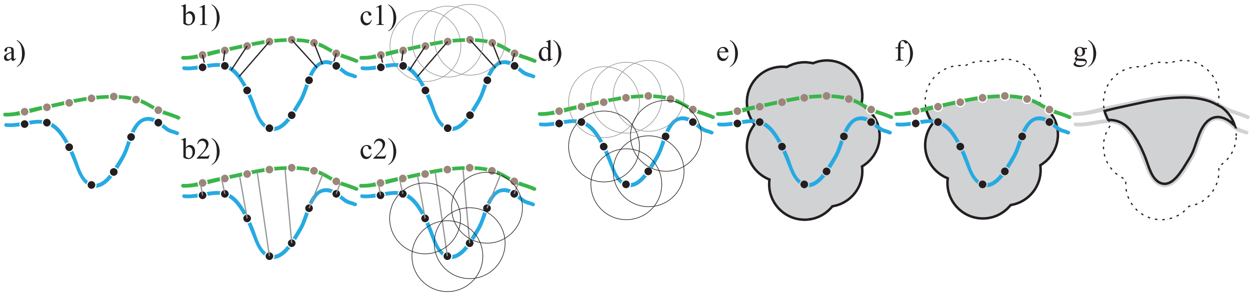


Fig. 4. Isolating large focusing regions. a) The focusing surface, which bounds the region below the grey line, and the molecular surface, which bounds the region below the black line, are composed of triangles, whose corners are illustrated with dots. b) We measure the distance (thin lines) between the points of the focusing surface and the closest position on the molecular surface (grey lines, b1) as well as the distance between the points of the molecular surface and the closest position on the focusing surface (black lines, b2). c) A sphere with radius 2 Å is centered on any point on either surface that is further than 2.0 Å from the other surface (dotted circles). d) The spheres are collected into a list and their Boolean union is computed (e). f) Next, we compute the the Boolean intersection (solid black line) between the focusing surface and the sphere union (dotted shape). g) Then we use the Boolean difference to remove the region within the molecular surface. The result is a non-thin region outside the molecular surface where the electrostatic potential is enhanced.

Next, we compute the Boolean union of all spheres (Figure 4e). With the resulting sphere union, we compute the Boolean intersection with the focusing surface, and the Boolean difference with the molecular surface. The result approximates the three dimensional regions where the molecular surface is distant from the focusing surface.

Inter-surface distances and sphere radii were calibrated on a two dimensional parameter sweep over all datasets, and manually inspected. We considered distances between 1.5 and 3 Å, at .5 Åincrements, and sphere radii between 1.5 and 3 were considered, at .5 Å increments. At smaller distances, spheres were generated too frequently, covering many thin focusing regions that are obviously unrelated to function. Larger distance thresholds failed to generate spheres in narrower cavities. Spheres having a radius of 1.5 Å tended to define focusing regions with empty zones inside, because they were too small to occupy the center of the cavity region. Sphere radii larger than approximately 2.5 Å generally appeared to create few differences, except that they added thin focusing regions on the boundary of a larger focusing region. On our data, setting distance thresholds at 2 Å and sphere radii at 2Å appeared to best isolate influential focusing regions.

3.4 A Statistical Model of Focusing Region Volume

After removing thin focusing regions, several regions can still be found on most proteins and DNA. More than 96% of these regions are small, non-thin areas that occupy less than 1 Å^3 in the crevices between surface residues. Functionally significant regions, on the other hand, can occupy hundreds of cubic angstroms. Since interactions with small regions could obviously occur by random chance, we designed a statistical model for filtering out such functionally insignificant focusing regions.

Our statistical model, based on a model described earlier [7], but paraphrased here, estimates the probability p that a given focusing region has *small* volume, and is thus functionally insignificant. Since a random selection of any focusing region is likely to result in a small region by random chance, we hold this position, our *null hypothesis*, by default. However, if p is lower than a given threshold of probability α (we use 0.001), then we discard the null hypothesis in favor of an *alternative hypothesis*: We assess that given region is *not small*, being in fact too large to occur by random chance, so it may thus be influential for biological function. These assessments are based on thousands of observations made on a training set, and not statements of fact.

To train our model, we randomly designated half of our DNA dataset as our training set. First, we identified every individual focusing region in the training set. Most structures exhibited approximately 100 individual regions, most smaller than one cubic angstrom. Together, a histogram of their volumes (figure 5) closely fits a log-normal distribution. The log-normal function, as an approximation of that histogram, is critical for estimating the probability p: Given the volume v of a query fragment, we estimate the probability p of observing a fragment with larger volume than v, at random. This probability is equivalent to estimating the volume under the curve to the right of v, relative to the total volume under the curve (see [7]), which is can be computed easily for the log-normal function. Comparing that probability to α, in the manner described above, enables the final prediction to be made: If $p > \alpha$, we say that the

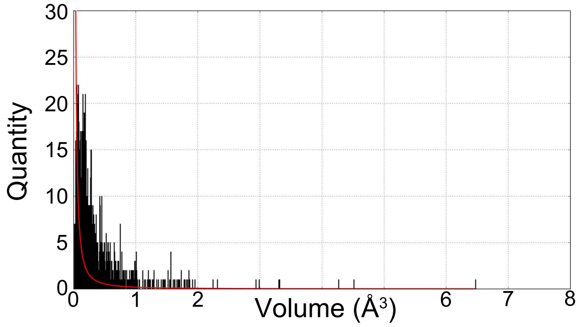


Fig. 5. Volume of non-thin focusing regions in DNA. The red line is the log-normal function fit to the data. Black bars represent the number of individual focusing regions at specific volumes. This graph illustrates approximately 99% of all training set data. The remaining 1% of individual focusing regions have much larger volumes in a long tail to the right, and are not shown, to permit a non-degenerate view of the data.

query fragment is statistically insignificant, and we assume the null hypothesis. If $p \leq \alpha$, we say the query is statistically significant, and accept the alternative hypothesis.

We selected α based on an inspection of large focusing regions in our training set. Focusing regions that exhibit bound arginines tend to be many tens or more than 100 cubic angstroms. A p-value greater than .001 corresponded to focusing regions larger than 47 cubic angstroms in our data set.

3.5 Comparing the focusing region

We use VASP [10] to compare the position of the focusing region in the minor grooves of DNA with the position of amino acids from proteins in complex with the DNA. As input, we begin with the region of focusing and the structure of an amino acid. First, we generate the molecular surface of the amino acid using the Trollbase library from GRASP2 [39]. Since molecular surfaces are closed surfaces, we can interpret them as three dimensional solids, enabling them to be compared with Boolean set operations. We then compute the Boolean intersection (e.g. Figure 1) between the region of focusing and the amino acid. The volume of the intersection region measures how much the intersection region and the amino acid coincide in space.

3.6 Data preparation

We use two data sets. The first is a case study on two proteins: Cu-Zn Superoxide Dismutase (pdb: 2SOD) and Trypsin (pdb: 3PTB). On SOD, atomic charges were assgined following the work of Klapper and coworkers [29]. Most notably His-41 was assigned a charge of +1 and His-61 was assigned -1 (charges were split between the ring nitrogens), as discussed here [15], [1], [55].

The second data set, provided by and studied first by Rohs and coworkers [46], is composed of 884 DNA-protein complexes. From this set, several structures were eliminated due to missing nucleotide side chains (1GJI, 1EGW, 1R71, 2W7N, 3L4J, 3L4K, 3N78, 3N7B and 3OD8), or lacking paired DNA chains (3HJF, 3HK2, 3HM9, 3HO1 and 3HVR). A final structure (2XSD) was removed because it caused an error in Delphi.

We removed ions, duplicated atoms, and waters from the remaining 866 structures. The *reduce* program (ver. 3.14) [61] from the MolProbity [12] package was then used to correct protonation. Finally, potential fields were computed using DelPhi 5.1 [44].

3.7 Implementation Details

Using the fields solved with Delphi, the method described in Section 3.1 generated the focusing region in an average of 3 seconds, using one core of an AMD Opteron 6128 with 2 gigabytes of random access memory (RAM) per core. Experiments were run on a system at Lehigh University called Corona, which contains 1040 such cores.

In-house software using the OpenGL library was used to generate the three dimensional visualizations in figures 3, 6, and 8 on Nvidia Geforce GT 330 hardware. Boolean set operations were used to generate the cross-sectional views in Figures 6 and 8.

The resolution parameter (.5 Å) was established by identifying a small decimal value that permitted our calculations to require less than 2 gigabytes of memory. .5 Å is considerably finer than the van der Waals diameter of heavy atoms and provides a realistic representation of the focusing region. The focusing threshold K was determined by visually observing 5 example regions where focusing is known to occur in DNA (such as in Figure 8). Sweeping values from .5 to 3.5 kT/e, in increments of .5, we selected 1 kT/e because the focusing region was large and other surface effects are minimized. 2 kT/e was used for SOD and trypsin, because it generated smaller focusing regions that could be easily visualized in cross section; other thresholds yielded related results.

4 Experimental Results

In this section, we first paraphrase our earlier results that demonstrate the accuracy of our methods for identifying regions of electrostatic focusing [11]. In sections 4.1 and 4.2, we show that our methods identify previously documented regions of electrostatic focusing in protein, on a small scale, and DNA, on a large scale. Our new results, in section 4.3, demonstrate how a volumetric analysis can be used to automatically eliminate thin, functionally irrelevant focusing regions, and how a statistical model can be

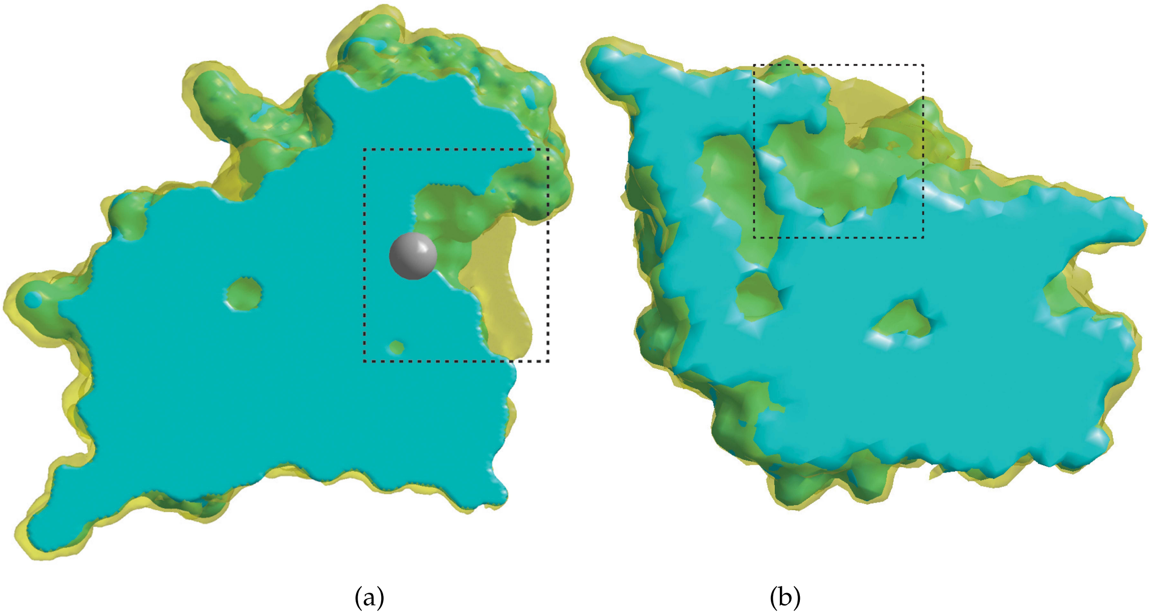


Fig. 6. A cross-section of the region of electrostatic focusing in SOD (a) and trypsin (b). In both panels, the molecular surface is shown in teal and the region of electrostatic focusing, where the uniform dielectric potential field and the nonuniform dielectric potential field differ by at least 2kt/e, is shown in transparent yellow. Binding regions, where considerable focusing occurs, are indicated by a dotted rectangle. SOD is shown with a copper ion cofactor, indicated by as a grey sphere.

used to further eliminate focusing regions that are so small as to have occurred by random chance.

4.1 Focusing Regions in Protein Structures

To demonstrate that Marching Cubes accurately detects instances of electrostatic focusing, we consider two proteins in which focusing is an established aspect of function. Superoxide dismutase (*SOD*) separates the toxic compound superoxide into hydrogen peroxide and oxygen. This reaction prevents superoxide from reacting with and damaging the larger molecular system and from forming other damaging compounds [2].

Superoxide molecules, which are negatively charged, have a strong affinity to the SOD binding site in part because of a their positively charged electric field. However, if we evaluate the field under a uniform dielectric model, the entire protein appears, from a distance, to be enclosed by negative potential [29]. Under those circumstances, SOD would not be attractive to its own substrates. A nonuniform dielectric model, used first by Klapper and coworkers [29], enhances potentials inside the cavity and makes the positive potential of the cavity apparent at longer distances.

Electrostatic focusing is also crucial for the activity of trypsin. As digestive proteins, trypsins hydrolyze peptide bonds following a positively charged amino acid [17]. This action is crucial for absorbing food proteins, which are too large to be absorbed in the small intestine otherwise. Trypsins employ a negatively charged binding pocket to stabilize positively charged sidechains for a crucial part of the hydrolysis reaction. However, estimates of the potential inside this binding pocket can vary considerably under a uniform dielectric model. This variability can be resolved by using a nonuniform dielectric, as observed by Soman and coworkers [54], producing plausible negative potentials that capture the focusing effect.

To demonstrate that our technique accurately identifies focusing regions, we examined the PDB structures of SOD (pdb: 2SOD) and trypsin (pdb: 3PTB). In both cases, we calculated the region where the potential is enhanced by more than 2 kT/e (i.e. $K > 2$). In SOD, the resulting region covered the molecular surface in a thin layer, less than 1 Å in depth. However, in the binding cavity, the focusing region was significantly deeper, nearly 8 Å. This trend is apparent in Figure 6a, which shows a cross section of SOD and its binding sites. In bovine trypsin (pdb: 3PTB), the focusing region again thinly covered many parts of the molecular surface, but entirely filled the arginine/lysine binding pocket to depths of over 10 Å (Figure 6b). These observations illustrate the accuracy of our method because they localize electrostatic focusing in the same regions as those identified by Klapper et al, in SOD, and Soman et al. in trypsin. Similar effects occur at other focusing thresholds ($K = \{1.5$ kT/e, 1.0 kT/e, 0.5 kT/e$\}$), in both SOD and trypsin.

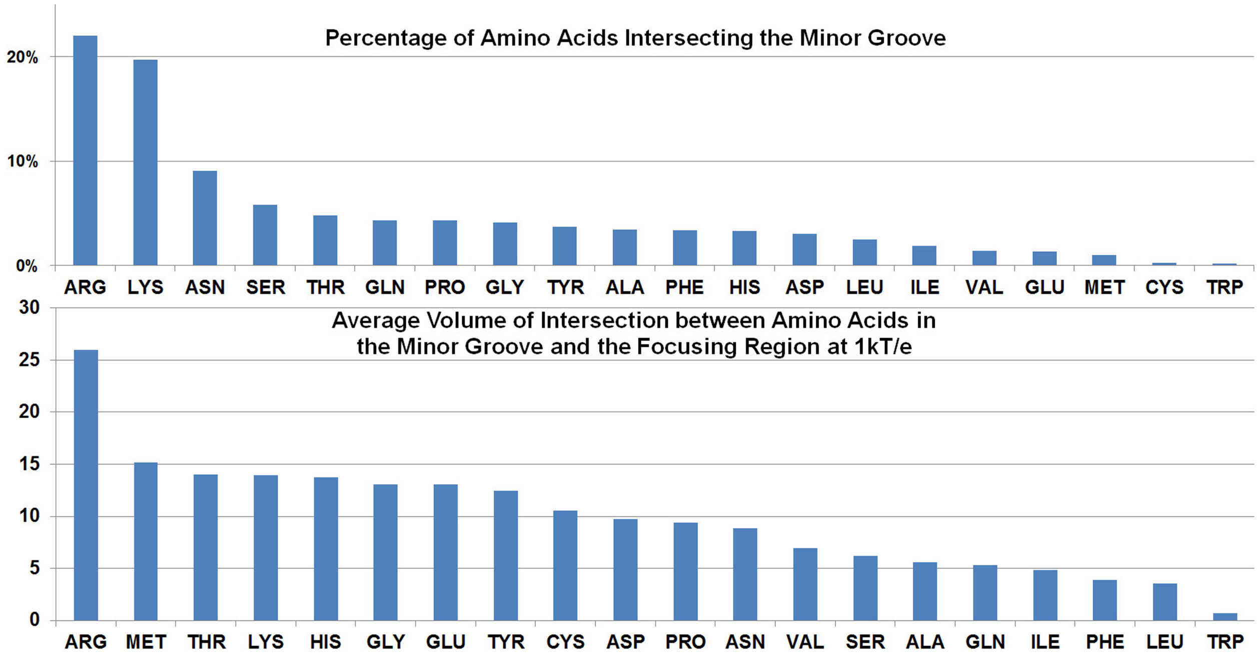


Fig. 7. A Survey of Amino Acid Interactions in the Minor Groove.

4.2 Focusing in Protein-DNA Complexes

DNA Shape can be exploited by proteins to selectively recognize regions of the double helix. Narrowness in the minor groove, for example, causes electrostatic focusing that enhances the negative potentials inside and promotes interactions with arginine sidechains. These observations, made by Rohs and coworkers [46] on a large number of protein-DNA complexes, revealed a mode of protein-DNA interaction that depends on electrostatic focusing.

We reproduced part of these observations to test the accuracy of our method for representing regions of electrostatic focusing. First, using the method described in section 3.2, we counted the number and sidechain of amino acids in the minor groove in our DNA-protein data set. These totals, shown in the top of Figure 7, reproduce the enrichment of arginines first observed by Rohs et al. [46]. Lysines, which also exhibit a distinctly positive net charge, were also enriched in the minor groove.

We then generated the focusing region (K = 1kT/e) in every DNA structure, and, as described in section 3.5, we computed the volume of intersection between the focusing region and every amino acid in the minor groove. The average volume of intersection for each type of amino acid is shown at the bottom of Figure 7. Arginines occupied an average of 25.9 Å^3 inside the focusing region, 72% greater than the second largest occupants, methionines (15.1 Å^3), and double or nearly double that of all other amino acids. These results reproduce the earlier observations that arginines are enriched in regions of electrostatic focusing [46], and while small differences in less-enriched amino acids were observed, they verify the accuracy

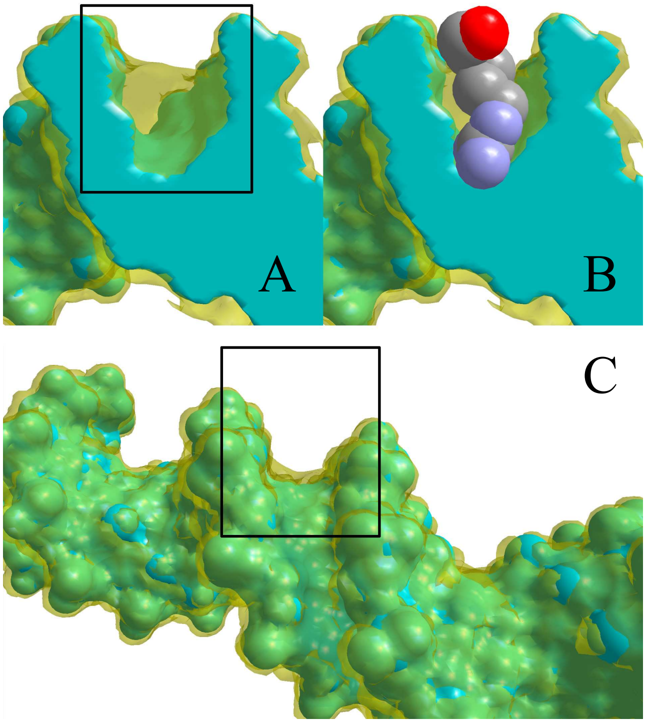


Fig. 8. Variations in electrostatic focusing in the minor groove. The molecular surface of DNA (pdb: 1HF0) is shown in teal and the surface of the focusing region at K = 1kT/e is shown in transparent yellow. The top two figures illustrate a cross section of the minor groove of DNA (black box) where electrostatic focusing is occurring, both without (A) and with (B) bound arginine. The bottom figure (C) illustrates a different region of the minor groove (black box) where less focusing is occurring.

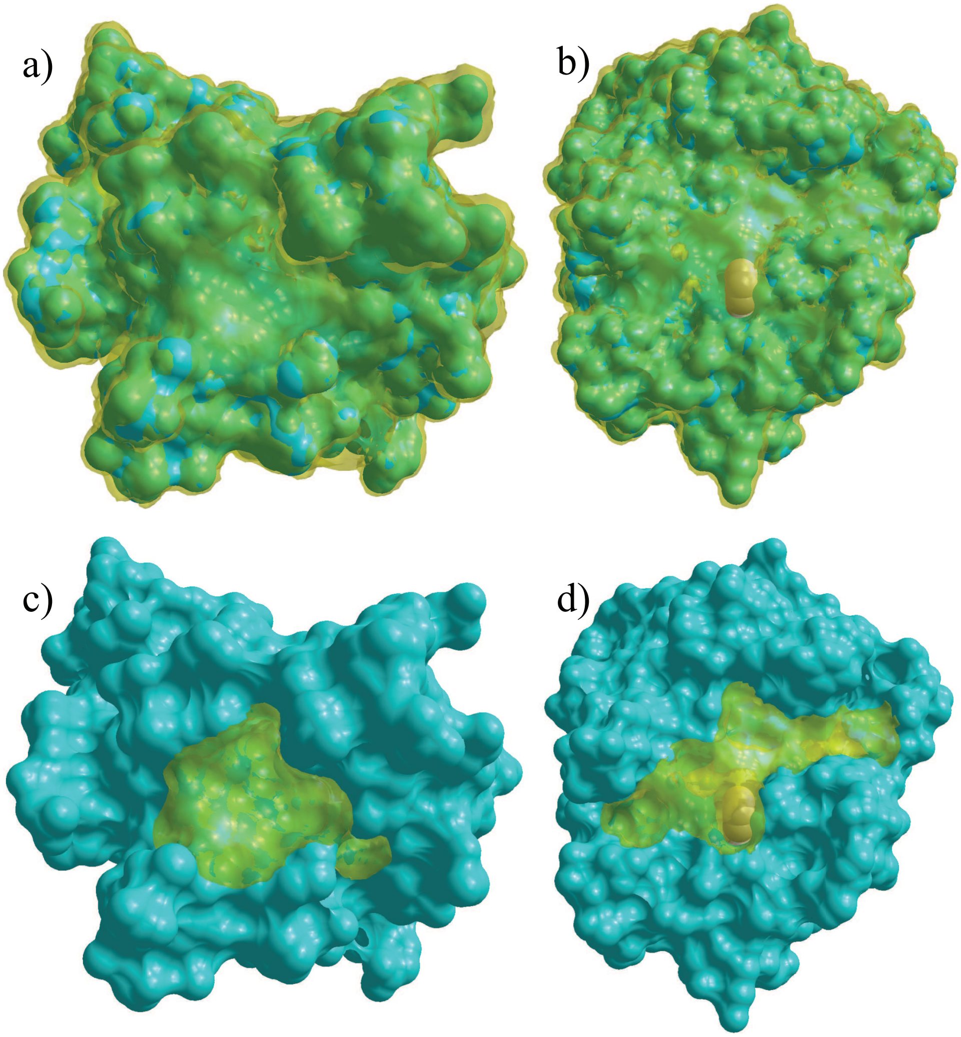


Fig. 9. Effect of Eliminating Thin Focusing Regions. The molecular surface of SOD (a, c) and trypsin (b, d) is shown in opaque teal. The region where electrostatic potential is enhanced by greater than 2 kT/e, in SOD (a) and trypsin (b), is shown in transparent yellow. After removing thin focusing regions, the largest remaining focusing region in SOD (c) and trypsin (d), as determined by the method described in Section 3.3, is drawn in transparent yellow.

of our representation.

Electrostatic focusing can vary considerably in the minor groove. In figure 8, close-up views of two different but nearby regions in the minor groove (from pdb 1HF0) illustrate this point. When present in the minor groove, arginine sidechains occupy the entire depth of the focusing region, completely immersing the guanidinium group in the focusing region (figure 8b). Ten base pairs away, where the minor groove is wider, the focusing region occupies only a thin film 1 Å in depth.

4.3 Isolating Influential Focusing Regions

The solid representations developed so far can effectively represent regions of electrostatic focusing, but they also contain small, flat regions where focusing is unlikely to influence biological function. These regions are distinctly visible in cross-section (Fig. 6) and from the exterior (Fig. 9a,b) of our protein data set and on DNA (Fig. 8c). To reduce flat, functionally irrelevant focusing regions, we developed a volumetric method for isolating larger regions, described in section 3.3.

On superoxide dismutase, 16 larger regions of electrostatic focusing were identified. The largest region, with a volume of 428 Å^3, was the functional site (Figure 9c). The remaining cavities were smaller, averaging 86 Å^3. Trypsin exhibited 24 non-flat regions of electrostatic focusing. The largest region occupied 765 Å^3, and in addition to the arginine/lysine binding cavity, it also occupied a considerable part of the peptide binding channel nearby (Figure 9d). The remaining regions were considerably smaller, averaging just 32 Å^3. While these results describe only two protein structures, they indicate that it is possible to distinguish functionally significant regions of electrostatic

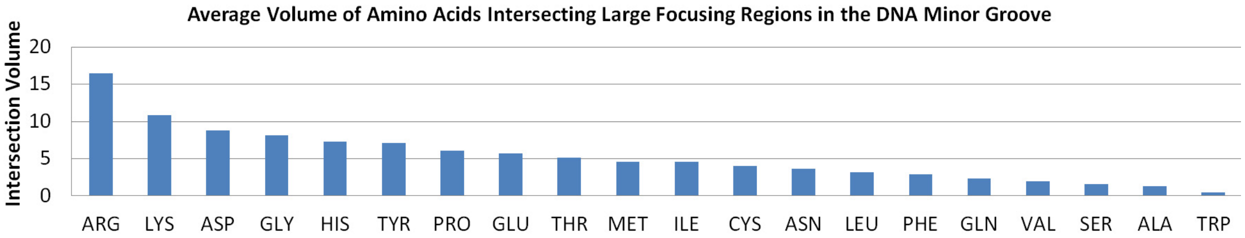


Fig. 10. Intersection volume, left, in cubic angstroms.

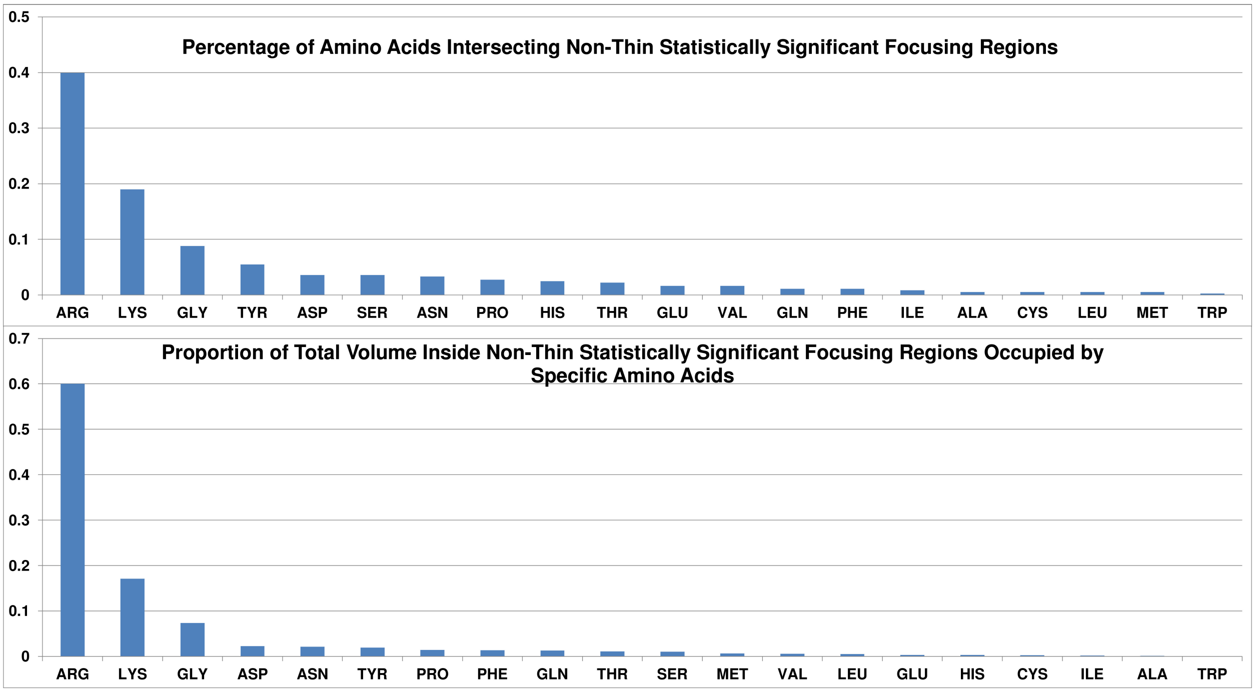


Fig. 11. A survey of amino acids intersecting focusing regions with statistically significant volume.

focusing from the many focusing regions that are too small to influence function. Figures 9 and 8c illustrate how completely the thin focusing regions cover the molecular surface of the proteins and DNA in our dataset, but our approach significantly narrowed the search for potential binding sites.

To evaluate our algorithm at a larger scale, we used the same parameters to identify large focusing regions on all DNA structures. If the focusing regions are correctly identified, then we expect average volumes of intersection with different amino acids in the minor groove, as defined earlier, to continue to strongly favor arginine. This effect can be seen in our data in Figure 10. The strong bias towards arginine remained despite the fact that the average volume of intersection dropped markedly for all amino acids when thin focusing regions were removed. Lysine, the second most common amino acid in the minor groove (see top of Figure 7), with the second most positive net charge, also had the second largest average volume of intersection. These results indicate that large focusing regions in the minor groove correctly identify regions that favor arginine binding.

The vast majority of focusing regions in the minor groove are smaller than 1 cubic angstrom. To further eliminate statistically insignificant amino acids, we trained our statistical model on one half of our DNA dataset, and estimated the p-values of focusing regions in the other half of the dataset. Eliminating any focusing region with a p-value greater than .001, we observed that the proportion of amino acids intersecting a statistically significant focusing region (Fig 11, top) was considerably enriched in arginines, even relative to our earlier findings in Figure 7. In fact, arginines occupied almost 60% of all volume occupied by any amino acid in a significant focusing region, and arginines and lysines together occupied nearly 80% (Fig 11, bottom). These results suggest amino acids other than arginine and lysine are almost always interacting with focusing regions that are so small as to occur by random chance.

5 Conclusions

This paper presents several new techniques for representing, comparing and analyzing regions where electrostatic focusing occurs. We verified the accuracy of our solid representations on trypsin and superoxide dismutase, which depend on electrostatic focusing to perform their biological function. We also verified the our representation at a large scale on 866 protein-DNA complexes, demonstrating that our software correctly reproduced earlier experiments that observed the enrichment of arginine sidechains in the minor groove. By comparing the position of the arginines with the solid focusing regions, our results indicate that our representation correctly detects regions where focusing occurs and enables the accurate and unsupervised comparison of focusing region geometry.

We also showed that large parts of the molecular surface are covered with thin focusing regions, while functionally influential focusing regions were deeper and occupied significantly more volume. To enable influential regions to be automatically detected, we described a volumetric algorithm for separating large focusing regions from the ubiquitous thin focusing regions. On superoxide dismutase and trypsin, the largest focusing region was always the binding site. In protein-DNA complexes, the large focusing regions were also enriched in arginine sidechains, as we observed earlier.

The largest focusing region on any molecular surface, however, is not a precise criterion for identifying functionally influential focusing regions. On some structures, no significant focusing regions exist, while DNA and other structures can exhibit several such regions. To filter out focusing regions that are too small, and assist in the prediction of functionally influential regions, we developed a statistical model of focusing region volume in protein-DNA complexes. Based on a significance threshold of .001, our model eliminates non-thin focusing regions that are too small to have occurred by random chance. The remaining regions were considerably more enriched in arginine sidechains than the focusing regions in our earlier data. These observations show that while the larger trend of arginine enrichment was still apparent, earlier observations included amino acids that co-occurred with focusing regions that were too small to be functionally relevant, and that the actual enrichment of arginines in the minor groove may be more selective than previously thought.

Our results indicate that volumetric algorithms and statistical models have useful applications in the analysis of electrostatic focusing. These techniques enable the dissection of electrostatic fields into discrete and functionally relevant focusing regions that can be isolated for further examination, pointing to new applications in protein engineering, where the design of electrostatic complementarity could enhanced, and in drug design, where statistically significant regions of electrostatic focusing could be used to consider new leads for selective binding.

ACKNOWLEDGMENT

The authors sincerely thank Remo Rohs for supplying the DNA data set and Debdas Paul for technical assistance.

REFERENCES

[1] S. A. Allison, G. Ganti, and J. A. McCammon. Simulation of the diffusion-controlled reaction between superoxide and superoxide dismutase. i. simple models. *Biopolymers*, 24(7):1323–1336, 1985.

[2] R. G. Alscher, N. Erturk, and L. S. Heath. Role of superoxide dismutases (sods) in controlling oxidative stress in plants. *Journal of Experimental Botany*, 53(372):1331–1341, 2002.

[3] 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*, 28(1):235–42, Jan. 2000.

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

[5] E. Catmull and R. Rom. *A class of local interpolating splines*. Academic Press, 1974.

[6] R. Chauvin. Explicit periodic trend of van der waals radii. *J Phys Chem*, 96:9194–7, 1992.

[7] B. Y. 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)*, pages 22–9, 2011.

[8] 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*, 14(6):791–816, 2007.

[9] B. Y. Chen, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge, and L. E. Kavraki. Algorithms for structural comparison and statistical analysis of 3D protein motifs. In *Pacific Symposium on Biocomputing.*, volume 345, pages 334–45, Jan. 2005.

[10] B. Y. Chen and B. Honig. VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity. *PLoS Comput Biol*, 6(8):11, 2010.

[11] B. Y. Chen and D. Paul. A Volumetric Method for Representing and Comparing Regions of Electrostatic Focusing in Molecular Structure. In *Proceedings of the 3rd ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB 2012)*, pages 242–249, 2012.

[12] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica D: Biological Crystallography*, D66:12–21, 2010.

[13] M. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–713, Aug. 1983.

[14] 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*, 406(5):713–729, 2011.

[15] E. D. Getzoff, J. A. Tainer, P. K. Weiner, P. A. Kollman, J. S. Richardson, and D. C. Richardson. Electrostatic recognition between superoxide and copper, zinc superoxide dismutase. *Nature*, 306, 1983.

[16] J. F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. *Curr Opin Struct Biol*, 6(3):377–85, June 1996.

[17] 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*, 85(14):4961–5, July 1988.

[18] S. C. Harvey. Treatment of Electrostatic Effects in Marcomolecular Modeling. *Proteins Struc Func and Genetics*, 5:78–92, 1989.

[19] Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein science : a publication of the Protein Society*, 3(2):211–26, Feb. 1994.

[20] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273(5275):595–603, Aug. 1996.

[21] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, 1995.

[22] B. H. Honig, W. L. Hubbell, and R. F. Flewelling. Electrostatic Interactions in Membranes and Proteins. *Annual Review of Biophysics and Biophysical Chemistry*, 15:163–193, 1986.

[23] W. A. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica A*, 34:827–828, 1978.

[24] A. Kahraman, R. J. Morris, R. a. Laskowski, and J. M. Thornton. Shape variation in protein binding pockets and their ligands. *J Mol Biol*, 368(1):283–301, Apr. 2007.

[25] E. Kangas and B. Tidor. Electrostatic Complementarity at Ligand Binding Sites: Application to Chorismate Mutase. *The Journal of Physical Chemistry B*, 105(4):880–888, Feb. 2001.

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

[27] K. Kinoshita and H. Nakamura. Protein informatics towards function identification. *Current Opinion in Structural Biology*, 13(3):396–400, June 2003.

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

[29] I. Klapper, R. Hagstrom, R. Fine, K. Sharp, and B. Honig. Focusing of electric fields in the active site of cu-zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *Proteins: Structure, Function, and Bioinformatics*, 1(1):47–59, 2004.

[30] R. A. Laskowski. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph*, 13(5):323–30, 307–8, Oct. 1995.

[31] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, 55(3):379–400, Feb. 1971.

[32] L. P. Lee and B. Tidor. Optimization of electrostatic binding free energy. *Journal of Medical Physics*, 106(21):8681–8690, 1997.

[33] W. E. Lorensen and H. E. Cline. Marching Cubes: A High Resolution 3D Surface Construction Algorithm. In *Proceedings of the 14th annual conference on Computer Graphics and Interactive Techniques (SIGGRAPH '87)*, volume 21, pages 163–170, 1987.

[34] J. B. Matthew. Electrostatic effects in proteins. *Annual review of biophysics and biophysical chemistry*, 14:387–417, Jan. 1985.

[35] S. McLaughlin. The electrostatic properties of membranes. *Annual review of biophysics and biophysical chemistry*, 18:113–36, Jan. 1989.

[36] H. Nakamura. Roles of electrostatic interaction in proteins. *Quarterly Reviews of Biophysics*, 29:1–90, 1996.

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

[38] E. K. O'Shea, R. Rutkowski, and P. S. Kim. Mechanism of specificity in the Fos-Jun oncoprotein heterodimer. *Cell*, 68(4):699–708, 1992.

[39] D. Petrey and B. Honig. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Method Enzymol*, 374(1991):492–509, Jan. 2003.

[40] F. Polticelli, B. Honig, P. Ascenzi, and M. Bolognesi. Structural determinants of trypsin affinity and specificity for cationic inhibitors. *Protein Science*, 8(12):2621–2629, 1999.

[41] J. W. Ponder and D. A. Case. Force fields for protein simulations. In V. Daggett, editor, *Protein Simulations*, volume 66 of *Advances in Protein Chemistry*, pages 27 – 85. Academic Press, 2003.

[42] M. T. Record, C. F. Anderson, and T. M. Lohman. Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity. *Quarterly Reviews of Biophysics*, 11(2):103–178, 1978.

[43] D. W. Ritchie and G. J. L. Kemp. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J Comput Chem*, 20(4):383, Mar. 1999.

[44] W. Rocchia, E. Alexov, and B. Honig. Extending the applicability of the nonlinear poisson-boltzmann equation: Multiple dielectric constants and multivalent ions. *Journal of Physical Chemistry B*, 105(28):6507–6514, 2001.

[45] N. K. Rogers. The modelling of electrostatic interactions in the function of globular proteins. *Progress in Biophysics and Molecular Biology*, 48(1):37–66, 1986.

[46] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig. The role of DNA shape in protein DNA recognition. *Nature*, 461:1248–1253, 2009.

[47] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng*, 11(4):263–77, Apr. 1998.

[48] R. B. Russell. Detection of protein three-dimensional sidechain patterns: new examples of convergent evolution. *J Mol Biol*, 279(5):1211–27, June 1998.

[49] S. Schmitt, D. Kuhn, and G. Klebe. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J Mol Biol*, 323(2):387–406, Oct. 2002.

[50] K. A. Sharp and B. Honig. Electrostatic Interactions in Macromolecules: Theory and Applications. *Annual Review of Biophysics and Biophysical Chemistry*, 19:301–332, 1990.

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

[52] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–47, Sept. 1998.

[53] C. V. Sindelar, Z. S. Hendsch, and B. Tidor. Effects of salt bridges on protein structure and design. *Protein Science*, 7(9):1898–1914, 1998.

[54] K. Soman, A. S. Yang, B. Honig, and R. Fletterick. Electrical potentials in trypsin isozymes. *Biochemistry*, 28(26):9918–9926, 1989.

[55] J. A. Tainer, E. D. Getzoff, K. M. Beem, J. S. Richardson, and D. C. Richardson. Determination and analysis of the 2 a-structure of copper, zinc superoxide dismutase. *J. Mol. Biol.*, 160(2):181–217, 1982.

[56] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

[57] C. D. Waldburger, J. F. Schildbach, and R. T. Sauer. Are buried salt bridges important for protein stability and conformational specificity? *Nature Structural Biology*, 2(2):122–128, 1995.

[58] L. Wang, T. O'Connell, A. Tropsha, and J. Hermans. Energetic decomposition of the alpha-helix-coil equilibrium of a dynamic model system. *Biopolymers*, 39(4):479–489, 1996.

[59] A. Warshel and S. T. Russell. Calculations of electrostatic interactions in biological systems and in solutions. *Quarterly Reviews of Biophysics*, 17(3):283–422, 1984.

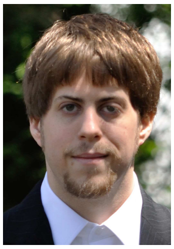
[60] W. C. Wimley, K. Gawrisch, T. P. Creamer, and S. H. White. Direct measurement of salt-bridge solvation energies using a peptide model system: implications for protein stability. *Proceedings of the National Academy of Sciences of the United States of America*, 93(7):2985–90, Apr. 1996.

[61] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Using hydrogen atom contacts in the choice of sidechain amide orientation. *Journal of Molecular Biology*, 285(4):1735–1747, 1999.

[62] 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*, 105(14):5441–6, Apr. 2008.

[63] A. S. Yang and B. Honig. Free energy determinants of secondary structure formation: I. alpha-Helices. *Journal of Molecular Biology*, 252(3):351–365, 1995.

[64] G. Zheng, X.-J. Lu, and O. WK. Web 3dna - a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Research*, 37:W240W246, 2009.



Seth Blumenthal received a B.S. degree in Bionengineering and a B.A. in Integrated Bioengineering and Computer Science from Lehigh University in 2012. He is currently working as a technical development consultant at Klish Group. Seth's previous research under Dr. Chen has included binding site analysis in Cytochrome P450 and the effect of protein flexibility in protein-ligand interactions. His research interests focus around developing software to automatically analyze and model volumetric protein data.

Wenjie Yang received the BS degree in computer science from South China University of Technology in 2011 and the MS degree in computer science from Lehigh University in 2013. He is currently working as a research assistant at the Informatics Lab at Lehigh University. His research interests include the application of machine learning on computational problems and structural bioinformatics.

Yisheng Tang received the BS degree in information security at Donghua University in 2011 and the MS degree in computer science from Lehigh University in 2013. He is currently working as a software developer at Quantum 3 media in New Jersey. His research interests include the application of statistical methods on bioinformatics and image processing problems.

Brian Y. Chen received the PhD degree in computer science from Rice University in 2007, where he was an NLM predoctoral fellow via the WM Keck Center for Computational and Structural Biology, and postdoctoral training at the Department of Biochemistry and Molecular Biophysics, the Center for Computational Biology and Bioinformatics, and the Howard Hughes Medical Institute at Columbia University. He is currently a P.C. Rossin Assistant Professor in the Department of Computer Science and Engineering at Lehigh University. His research is focused on computational techniques that reveal the determinants of binding specificity in protein structures. His work encompasses concepts from structural biology, computational geometry, nonparametric statistics, and high performance computing. He is a member of the IEEE, the ACM, and the ISCB.