Conformational sampling reveals amino acids with a steric influence on specificity

Brian Y. Chen and Ziyi Guo
Dept. Computer Science and Engineering
Lehigh University
chen@cse.lehigh.edu

May 22, 2015

Abstract

Flexible representations of protein structures can enable structure comparison algorithms to find remotely homologous proteins, even when they have been crystallized in different conformations. By compensating for large spatial variations, these representations can enable these algorithms to better detect remote similarities in the space of protein structures. Subtle variations in protein structures can also have a substantial impact structure comparison. For example, the motion of a single side chain into a binding cavity can make the cavity appear totally dissimilar to identical binding sites, even though, in reality, the presence of the side chain does not affect binding. To address the impact of subtle conformational variations, this paper describes FAVA (Flexible Aggregate Volumetric Analysis), an algorithm that enables comparisons of ligand binding sites while compensating for subtle, localized flexibility.

FAVA integrates hundreds of conformational samples, sourced from any molecular simulation software that provides all-atom detail, to characterize the geometry of ligand binding sites as they frequently appear. This representation enables rare conformations, as defined by the user, to be excluded from the structural comparison. In our results, on three families of serine proteases and three families of enolases, we show that despite substantial binding site variations, FAVA is able to correctly classify families with different binding preferences. We also demonstrate that FAVA can examine the motion of individual amino acids to identify those that influence ligand binding specificity. Together, these capabilities demonstrate that comparison errors associated with small conformational variations, which can substantially alter the geometry of ligand binding sites and other local features, can be mitigated by an analysis of many conformational samples.

1 Introduction

Algorithms that compare protein structures have generally focused on the identification of proteins with similar functions at great evolutionary distances. Aiming to reveal the topology of the space of protein structures [35, 38, 25, 57], many algorithms specialize in identifying proteins with backbone similarities that can point to shared evolutionary origins even when similarity is not apparent from amino acid sequences alone [37, 39, 26, 43, 49, 59]. In such cases, proteins that have been crystalized in different conformations can be overlooked, because the spatial relationships between their constituent parts can appear quite different. To compensate for conformational variations, some methods employ flexible representations of proteins as collections of rigid components with flexible linkers. These representations employ hinges [47, 20], graph structures [58, 29], fragments [34], and dynamic programming [56, 7, 32] to avoid discarding similar proteins in alternate conformations.

Comparisons of smaller elements of protein structure, such as ligand binding cavities [43, 6, 10, 8], are also affected by small conformational variations. While large changes, such as backbone motions, still disrupt the comparison of ligand binding cavities [17, 18], smaller variations, such as the motion of individual side chains, also cause similar binding cavities to appear very different. This second source of error disrupts the search for proteins with similar binding sites at great evolutionary distances [11, 54, 15, 8], and also the

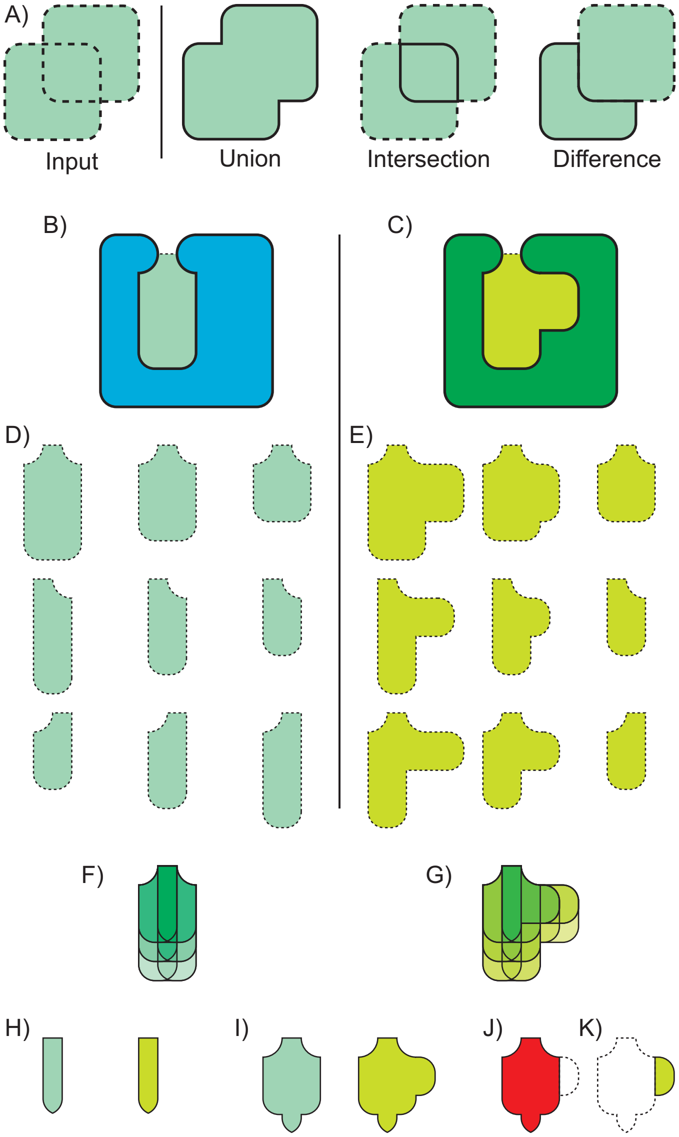

Figure 1: An overview of the FAVA method. A) CSG operations used by FAVA, with input regions (green, dotted outline) and output regions (solid outline). B,C) Input proteins X (blue) and Y (green) with ligand binding sites x (light green) and y (yellow). D,E) x and y in different conformational samples. F,G) All conformational samples of x (transparent green) and y (transparent yellow), superposed, with black outlines. Considerable variations in cavity shape are apparent. H) Using CSG, the intersection of all cavity regions in both proteins is too small to accommodate ligands. They are also identical, revealing little about different binding preferences. I) Using CSG, FAVA approximates frequent regions, where every point is inside at least two thirds of all samples of x (green) and y (yellow). J) The intersection of frequent regions indicates regions that might accommodate similar molecular fragments (red). K) The difference between frequent regions (yellow) indicates a region where Y might often accommodate a ligand that X cannot, causing a difference in specificity.

identification of binding site variations that cause differences in ligand binding specificity [12, 22, 21]. While flexible representations can compensate for large variations in backbone geometry, existing methods do not compensate for the large space of smaller motions that disrupt local comparisons.

To address this problem, we describe a method for reducing comparison errors due to small variations in protein structure called FAVA (Flexible Aggregate Volumetric Analysis). FAVA integrates structural data from many conformational samples to create a volumetric representation of binding cavity regions (*frequent regions*) that are frequently accessible to solvent. In our results we will show that comparisons of frequent regions can reveal differences between binding cavities that explain different binding preferences, even though many conformational variations can obscure this trend. We will also demonstrate that comparisons of many conformations of individual amino acids with the binding cavities of other proteins can precisely reveal amino acids that have a steric influence on ligand binding specificity, even though many amino acids occasionally interfere with binding cavity shape. FAVA demonstrates that an aggregate analysis of many conformational samples can yield insights into conformational trends for localized comparison.

Sampled conformations exhibit a novel advantage. All-atom samples provide a more detailed representation of molecular motion than that provided by existing methods. In each sample, the position of every atom is affected by energy functions that enforce biophysical constraints. Each relevant atom can thus be used for constructing a semi-realistic representation of binding site geometry. In constrast, existing methods approximate rigidity in some places and flexibility in others. While the same biophysical constraints are not in effect, the computational expense of molecular simulation is also not necessary in existing methods. Detailed biophysical semi-realism is thus achieved in exchange for the substantial computational cost of conformational sampling. This tradeoff specializes FAVA for the detailed comparison of ligand binding cavities, where the motion of every atom is critical, whereas the detection of similar proteins with large conformational variations, which would be expensive to simulate anyway, is best suited for the rigid components and flexible linkers used in existing work.

Frequent regions are generated using boolean *operations* from Constructive Solid Geometry (CSG) [12], such as union, intersection and difference (Fig. 1a). CSG operations are also used to compare binding cavities and identify potentially influential amino acids. CSG intersections between frequent regions identify *conserved frequent regions*, which describe regions in the binding cavities of two structurally aligned proteins that are frequently solvent accessible in both proteins. These regions might accommodate molecular fragments common to substrates that bind both proteins. CSG differences between frequent regions identify *unconserved frequent regions*, which represent regions in aligned binding cavities that are frequently solvent accessible in one protein but not the other. Unconserved frequent regions might therefore accommodate ligands in one protein that cannot bind in another and thereby have a steric influence on specificity. Finally, CSG operations permit the identification of amino acids that frequently alter cavity shape, and thus have a steric influence on specificity. Together, these techniques create a conformationally general approach for examining closely related proteins in search of influences on binding specificity.

In earlier work [22], we demonstrated that comparisons of frequent regions generated with FAVA could reveal patterns of subtle differences in binding cavities that relate to ligand binding preferences. We also showed examples demonstrating that amino acids that frequently alter cavity shape have a steric influence on specificity. This paper summarizes the methods employed by FAVA and extends our earlier examples, providing a comprehensive study of the conformational variation of all amino acids within a sequentially nonredundant set of serine proteases.

2 Methods

Below, we paraphase methods associated with FAVA that we described earlier [22]. FAVA defines frequent regions as regions in space that are solvent accessible in more than k/N samples, where k, the *overlap threshold*, is provided as input, and N is the number of samples. By setting k, the user defines how frequently a region in space must be solvent accessible in order to be part of the frequent region. Unusual gaps and clefts that occur less frequently than k/N will thus not be reflected in the shape of the frequent region. We first describe how frequent regions are computed using CSG operations, employing the symbols $\cup$, $\cap$, and $-$ to represent CSG unions, intersections and differences, respectively. Implementation of individual CSG operations has been described earlier [12]. Next, we describe how frequent regions from multiple proteins are compared using CSG operations. Finally, we explain how we use sampled conformations of individual amino acids to assess their steric impingement on nearby binding cavities.

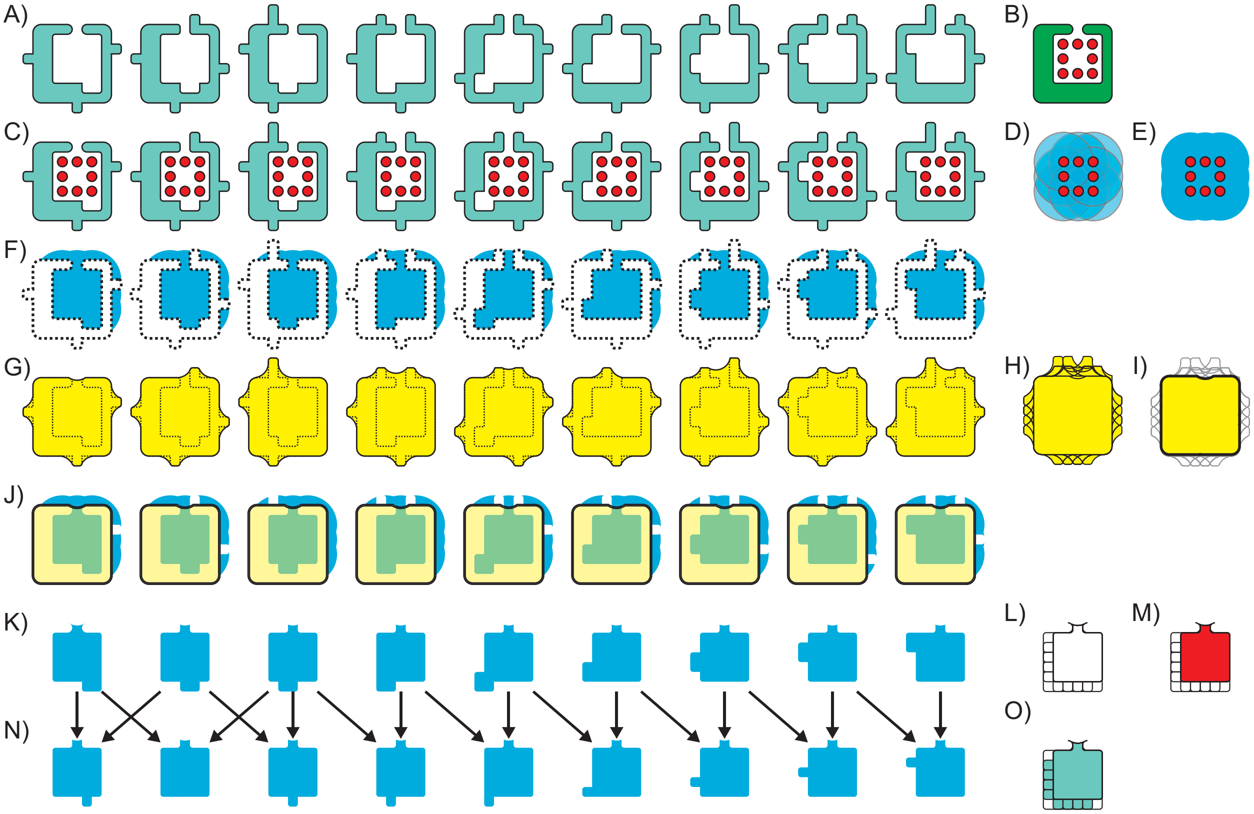

Figure 2: Generating frequent regions. A) Conformational samples $A_0, A_1, \ldots A_N$ of protein A, shown as molecular surfaces (teal, black outlines). B) Ligand l (red dots) bound to protein A (green). C) Aligning each A_i to A permits l to mark the ligand binding site in A_i. D) Spheres that define the neighborhood around the atoms of l (transparent blue, black outlines). E) CSG union of the spheres, S_l (blue). F) CSG difference A_i' that removes the region within the molecular surface of each A_i (dotted outline) from copies of S_l. G) Envelope surfaces (yellow, black lines) and molecular surfaces (dotted lines) of all A_i. H) Envelope surfaces aligned. Outlines of all envelopes are shown in black. I) The global envelope region, $E(A)$, generated with CSG intersections (yellow, heavy black outline). J) CSG intersection between each A_i', shown in blue, and $E(A)$ (transparent yellow, black outline). K) Cavities a_i defined on each conformational sample. L) Cavity borders superposed (black outlines). M) frequent region that overlaps at least 3 cavities (red). N) CSG intersections between several pairs of a_i. O) The CSG union of intersections in N: $\alpha_k^\star$, the approximated frequent region (teal) overlapping at least 2 cavities.

2.1 Computing frequent regions

To generate a frequent region, we begin with the overlap threshold k, N conformational samples of a protein structure A, and a ligand l bound to the structure of A (Fig. 2b). We use $A_0, A_1, \ldots A_N$ to refer to conformational samples of A, which could be provided from any source that specifies all atom positions. It is asssumed that enough samples have been gathered so that the range of short timescale motions, such as sidechain movements, are represented in the samples. Beginning with this data, we follow two steps to generate frequent regions: First, we define the ligand binding cavity in every sample, and second, we analyze these cavities to determine the shape of the frequent region.

We begin by superimposing every sample A_i onto A by computing a backbone superposition. Sophisticated algorithms designed to detect remote homologs (e.g. [37, 39, 26, 49, 59]) are not necessary, because the proteins in A_i are identical to A, except in their conformation. For this reason, we compute a superposition that minimizes the root mean squared distance between identical atoms [55]. Next, we generate the molecular surface $m(A_i)$ of every conformational sample A_i, using GRASP2 [41] (Fig. 2a). GRASP2 uses

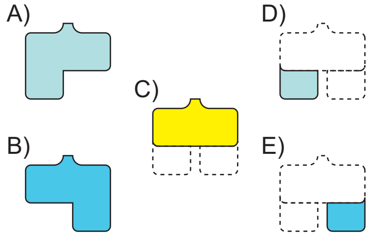

Figure 3: A comparison of frequent regions. A,B) Frequent regions $\alpha_k^\star$ (teal) and $\beta_k^\star$ (light blue). C) Conserved frequent region, $FC(A, B)$ (yellow). D,E) unconserved frequent regions (teal, light blue).

the classical rolling probe algorithm [13] and the standard probe size of 1.4Å.

Now, since every A_i has been superposed onto A, the binding cavity of A is now approximately superposed with itself in every sample, so we use the ligand l, as bound in A, to represent the neighborhood of the binding site in every surface $m(A_i)$ (Fig. 2c). Specifically, we define a sphere of radius 5 Å , centered at every atom in l (Fig. 2d), and compute the CSG union of all these spheres. This neighborhood, S_l (Fig. 2e), defines the vicinity of the ligand binding cavity in every sample.

Next, we use S_l to define the binding cavity in every sample. First, we make a copy of S_l for every A_i and compute the CSG difference, $A'_i = S_l - m(A_i)$, revealing part of the cavity 2f). We also make an *envelope surface*, $e(A_i)$, for every sample. Here, we again use GRASP2, which uses the same rolling probe method except that the probe radius is changed to 5.0Å (Fig. 2g). This larger probe is 10 Å in diameter so it does not roll into small clefts and cavities. For this reason, we use the resulting surface $e(A_i)$ to define a logical exterior boundary between the cavity and the solvent. Since $e(A_i)$ can vary significantly between different conformational samples, especially because of solvent-facing side chains (Fig. 2g) unrelated to cavity shape, we mitigate these differences by computing their intersection (Fig. 2h,i),

$$E(A) = \bigcap_{\forall i} e(A_i). \tag{1}$$

As the intersection of the envelopes from all conformational samples, we refer to $E(A)$ as the *global envelope region*. Note here that because our samples are generated at a medium timescale, and thus exclude large backbone motions, the shape of $E(A)$ is not strongly influenced by backbone motion. Next, for every i, we compute the CSG intersection of A'_i and $E(A)$ (Fig. 2j). The result is the binding cavity in every conformational sample, which we call a_i (Fig. 2k).

Once we have determined the shape of the binding cavity in each conformational sample, we then use all a_i to approximate the frequent region α_k. Here, it is crucial to note that α_k is not practical to compute explicitly on a protein with many sampled conformations. Consider, for example, the simple case of $k = 30$. The region α_{30} includes the CSG intersection of $a_0, a_1, a_2, \ldots, a_{30}$, because any point inside all of these regions is inside at least 30 a_i, and thus inside α_{30}. The same is true for any thirty member subset of $\{a_0, a_1, \ldots, a_N\}$, so α_{30} is the union of all intersections of thirty distinct sample cavities: $\binom{N}{30}$ intersections. Where N is several hundred samples and k is nontrivial, the exponential size of the calculation is clearly impractical, given the number of combinations.

Even though an explicit computation of α_k would be impractical, it can still be rapidly approximated. For any k, we select 500 random and distinct subsets of $\{a_0, a_1, \ldots, a_N\}$ of size k. Next, we compute the intersection of the binding cavities in each subset, and then the union of all such intersections. We call this region $\alpha_k^\star$. Fig. 2n illustrates a small example of this process on a random selection). We tried random selections of different sizes [22] and found that frequent regions approximated from random subsets of 500 had consistent volumes, and therefore deemed 500 sampled subsets to be sufficient for accurate representations.

2.2 Comparing frequent regions

To compare frequent regions representing the binding cavities of two proteins, A and B, we begin by aligning B onto A using ska [57]. Next, we superpose every snapshot A_i onto A and every snapshot of B_i onto B by minimizing the root mean squared distance between identical atoms [55]. If more than two proteins were being considered, they would also be aligned to A first. Ultimately, every snapshot is aligned to A and B, which were first aligned onto each other. Using these alignments, we compute the frequent regions of A and B, using the method above.

Once the frequent regions $\alpha_k^\star$ and $\beta_k^\star$ are computed, we use $\alpha_k^\star$ and $\beta_k^\star$ to compute conserved frequent regions. We define the conserved frequent region $FC(A,B) = \alpha_k^\star \cap \beta_k^\star$ (Fig. 3c). We use the intersection because it approximates a region in both binding cavities that is solvent accessible in both proteins for more than k conformational samples. We quantify the difference between $\alpha_k^\star$ and $\beta_k^\star$ using the *volumetric distance*

$$D(A,B) = 1 - \frac{|FC(A,B)|}{|\alpha_k^\star \cup \beta_k^\star|}, \tag{2}$$

where $|x|$ denotes the volume inside region x. We measure volume using the Surveyor's formula [44], which we described earlier [12].

Unconserved frequent regions are treated in a similar manner. With the samples of A and B, we refer to the unconserved frequent region as $FV(A,B) = \alpha_k^\star - \beta_k^\star$ (Fig. 3f,g), and quantify the difference as $|FV(A,B)|$.

Finally, a comparison of cavities a_i and b_j from individual conformational samples of two different proteins is also possible. We evaluate their volumetric distance as

$$d(a_i,b_j) = 1 - \frac{|a_i \cap b_j|}{|a_i \cup b_j|}. \tag{3}$$

2.3 Influential amino acids

Given two proteins A and B, if the cavity of A is frequently different from B, then some set of amino acids is responsible for making these cavities different on a frequent basis. We identify such amino acids with FAVA.

At the level of individual samples, consider two samples of A and B, called A_i and B_j, and an amino acid r in A. We say that r makes the cavity a_i different from the cavity b_j if the intersection of the molecular surface of r in A_i, called $m(r_i)$, has a nonempty intersection with b_j. If so, then $m(r_i)$ occupies a region that is not solvent accessible in a_i but solvent accessible in b_j. Between these two samples r_i is thus one cause for the difference between a_i and b_j.

To evaluate how frequently r, an amino acid of A, creates differences between the cavities of A and B, we compute $INT_r(A,B)$, the median volume of intersection $|m(r_i) \cap b_j|$, for all pairs of samples A_i and B_j. When $INT_r(A,B)$ is large, then r frequently makes the cavity of A different from B; small values indicate that it rarely does.

2.4 Data set construction

The purpose of our experimentation is to demonstrate that FAVA can distinguish proteins with different binding preferences, even though conformational flexibility can make binding sites with different binding preferences seem erroneously similar. To test this hypothesis, we selected two superfamilies of proteins, the serine proteases and the enolases. Within each superfamily, we selected three distinct families with different binding preferences (Figure 2.4). Among the serine proteases, we selected the trypsin, chymotrypsin, and elastase subfamilies, and among the enolases, we selected the enolase, mandelate racemase, and muconate lactonizing enzyme families.

Serine proteases selectively cleave peptide bonds using a nucleophilic serine residue. Preferences for hydrolyzing a specific scissile bond are achieved by recognizing amino acids on both sides of the bond, most notably the $P1$ residue immediately before the bond. The $S1$ specificity pocket, which recognizes $P1$, is large and hydrophobic in chymotrypsins and prefers to bind large hydrophobic residues [33]. In trypsins, $S1$ stabilizes positively charged amino acids, complementing its notable negative charge [19]. Enolases exhibit a small hydrophobic $S1$ cavity that binds small hydrophobic amino acids [4].

Members of the enolase superfamily exhibit a TIM-barrel fold and an N-terminal "capping domain" [42]. Using amino acids at the C-terminal ends of beta sheets in the TIM-barrel, superfamily members achieve a range of different functions that generally abstract a proton from a carbon adjacent to a carboxylic acid [1]. The enolase family catalyzes the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate [30], mandelate racemases convert (R)-mandelate to and from (S)-mandelate [45], and muconate lactonizing enzyme catalyze the reciprocal cycloisomerization of cis,cis-muconate and muconolactone.

Serine Protease Superfamily:
Chymotrypsins: 1ex3
Elastases: 1b0e, 1elt
Trypsins: 1a0j, 1ane, 1aq7, 1bzx, 1fn8, 1h4w, 1trn, 2eek, 2f91
Enolase Superfamily:
Enolases: 1iyx, 1ebh, 1te6, 3otr
Mandelate Racemase: 1mdr, 2ox4
Muconate Lactonizing Enzyme: 2pgw

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

The structures of serine protease and enolase proteins were selected from the Protein Data Bank (PDB) [5] on 6.21.2011. Using Enzyme Commission classifications (EC number), we found 676 serine protease and 66 enolase structures among the families selected for our data set. From this group of structures, proteins with mutations, disordered regions, or enolases in closed or partially closed (and thus inactive) conformations were removed. From the remaining set, a set of sequentially nonredundant representatives were selected such that no representative had greater than 90% sequence identity to any other representative. Technical problems with simulation prevented 8gch, 1aks, and 2zad from being included in this set. From each of the remaining structures we removed waters, ions, hydrogens, and other non-protein atoms.

We superposed all serine protease sterutures against bovine chymotrypsin (pdb: 8gch), and all enolases against mandelate racemase from pseudomonas putida (pdb: 1mdr). We selected these structures because they were crystallized in complex with a ligand. As mentioned in Section 2.1, we use the position of the ligand to define the binding cavity in all conformational samples.

2.5 Gathering sampled conformations

FAVA is agnostic to the source of conformational samples, as long as each conformational sample is provided with full atomic detail. Naturally, the accuracy of FAVA relies on the accuracy of the underlying sampling technique, and we acknowledge the ongoing debate on the topic [27], but many techniques, including GROMAS [24], used here, have demonstrated considerable successes for some time. With unlimited computational resources, we would have examined multiple sources of conformational samples, but such a study was impractical with available tools.

For each data set structure, we computed conformational samples using GROMACS 4.5.4 [24]. Before the simulation, a cubic waterbox was created, and the protein was centered in the box. Waterbox size was set to contain the solute protein structure and a 1.0 nanometer margin between the protein and the nearest point on the box. Next, the waterbox was populated using SPC/E, an equilibrated 3-point solvent model [3]. Throughout the equilibration and simulation steps, fully periodic boundary conditions were used. Charge balanced sodium and potassium ions were then added to the solvent at a low concentration ($< 0.1\%$ salinity).

We then performed an energy minimization of the entire system using a steepest descent algorithm. In four 250 picosecond steps, we performed isothermal-isobaric (NPT) equilibrations to allow the solvent to equilibrate temperature and pressure prior to the primary simulation. Starting at 1000 $kJ/(mol * nm)$, each step reduced the position restraint force by 250 $kJ/(mol * nm)$ over the 1 nanosecond minimization period. Backbone position restraints were released for the primary NPT simulation.

System energies were generated at the start of the equilibration phase. Initial temperature was 300 Kelvin and initial pressure was 1 bar. The Nosé-Hoover thermostat [3] was used for temperature coupling. The P-LINCS [23] bond constraint algorithm was used to update bonds. Electrostatic interaction energies were calculated by particle mesh Ewald summation (PME) [24]. The Parrinello-Rahman algorithm was used for pressure coupling [40, 36]. All temperature and pressure scaling was performed isotropically.

Finally, the atomic positions and velocities of the final equilibration state were used to start the primary simulation. The simulation was sustained for 100 nanoseconds, in 1 femtosecond timesteps. P-LINCS and PME were chosen for their parallel efficiency. OpenMPI was used for inter-process and network communication. Simulations were run on multiple nodes with 16 cores each, with PME distribution automatically selected by GROMACS.

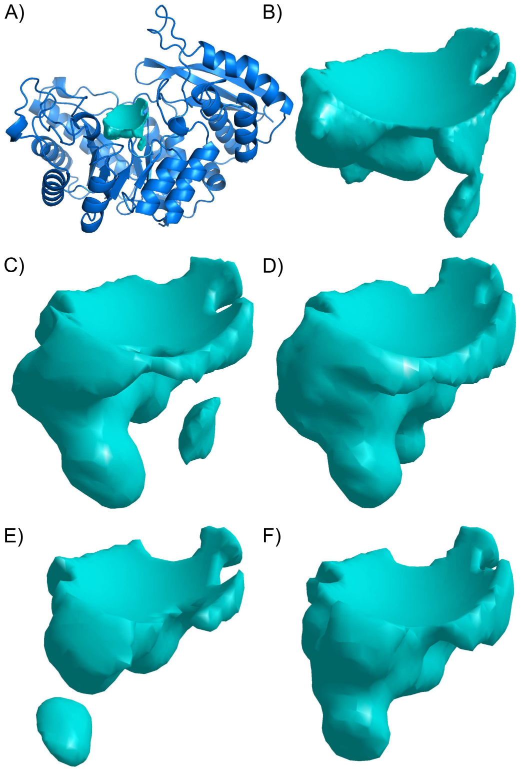

Figure 5: Conformational samples of the ligand binding cavity in yeast enolase (pdb: 1ebh). A) The position of the cavity (teal) within the tertiary structure of enolase (blue cartoon). B) The ligand binding cavity in the original crystal structure. C-F) Binding cavities from other conformational samples of yeast enolase. All panels illustrate the cavity from the same perspective, generated with the global envelope surface, as described earlier.

After completing each simulation, the trajectory file was converted into individual timesteps in the PDB file format, with waterbox atoms removed. Each timestep was then rigidly superposed onto the original structure by minimizing RMSD between identical atoms [55]. From these timesteps, we selected 600 conformational samples at uniform intervals, and used them to compute frequent regions.

2.6 Clustering frequent regions by volumetric distance

Once the volumetric distance between two frequent regions can be measured, using the method described in Section 2.2, we can visualize the pattern of similarities and differences between multiple frequent regions using a clustering algorithm. To perform this visualization, we first generated frequent regions from our dataset with an overlap threshold of 50. Then, we measured volumetric distance between all pairs of frequent regions in the same superfamily. Finally, we used the neighbor tool from the Phylogeny inference package Phylip [16] to perform UPGMA clustering (Unweighted Pair Group Method with Arithmetic mean) [51] based on volumetric distance.

To evaluate how well frequent regions avoid inaccuracies that may be caused by unusual conformational samples, we also computed 10 UPGMA clusterings of binding cavities from individual conformational samples selected randomly from each simulation. We visualized these all clusterings using Newick Utilities version 1.6 [28].

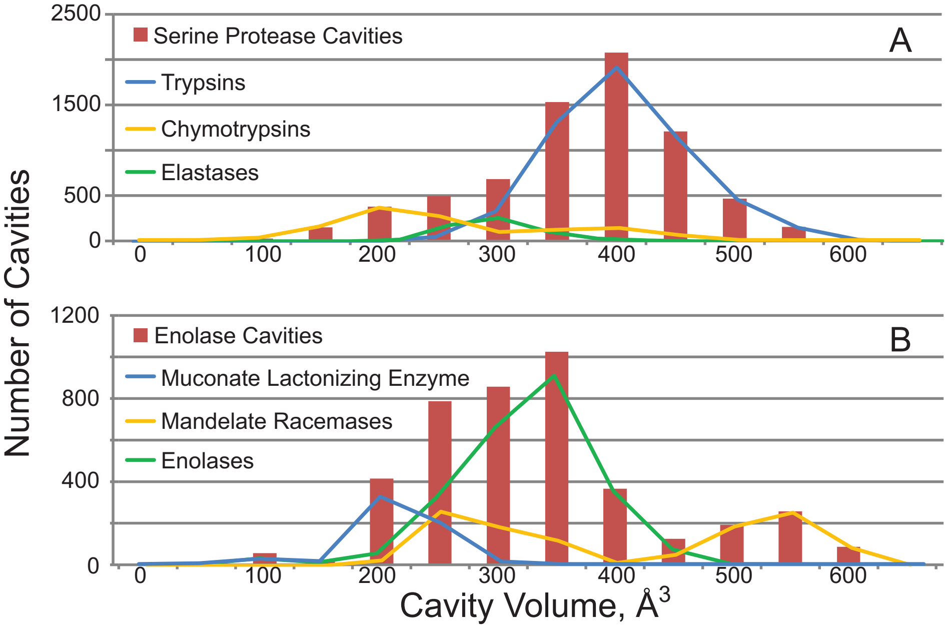

Figure 6: Volume of cavity sizes observed in conformational samples of serine proteases (A) and enolases (B), shown in red bars. Lines plot the quantity and spatial volume of cavities sampled for specific families.

2.7 Comparing FAVA against statistical models for rigid comparison

In earlier work, we developed VASP-S [9], a statistical model of volumetric variations between ligand binding cavities with identical binding preferences. Once trained, VASP-S can identify variations in binding cavity shape that are too large to be consistent with identical specificity, indicating that the cavities compared exhibit different binding preferences. We hypothesize that the variations that occur between different conformational samples from the same binding cavity are so substantial that VASP-E would incorrectly label some conformational samples as having different binding preferences, while FAVA would not make the same mistakes. These predictions would demonstrate that FAVA enables accurate comparisons of binding cavities in the flexible context.

2.8 Implementation Details

FAVA is a high-level procedure that uses CSG operations from VASP [12]. Running time is proportional to the volume of the inputs. CSG operations involving amino acids and binding cavities complete in fractions of a second, while operations on molecular surfaces require approximately 10 seconds. All computations were run on a cluster of 16 core machines equipped with AMD Opteron 6128 processors and 32 gigabytes of system memory. Figure 5 was generated with custom software and Pymol [14].

3 Results

3.1 Sampled conformations vary significantly

Figure 5 illustrates conformational samples of the ligand binding cavity of yeast enolase (pdb: 1ebh). Each sample was aligned against the original PDB structure, and is shown from the same orientation. It is evident that smaller backbone and sidechain motions created substantial variations that changed the geometry of the cavity. The degree of variation exhibited in Figure 5 was not the most variable, relative to other cavities in the data set, nor was it the most conserved. Binding cavities in some proteins, such as atlantic salmon pancreatic elastase (pdb: 1elt), varied much more, while others varied less.

To further illustrate the degree of structural variation between ligand binding cavities, we plotted the

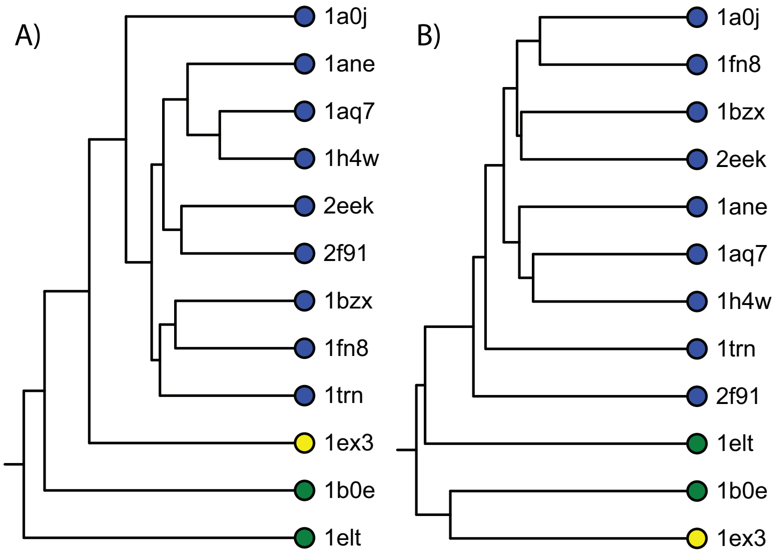

Figure 7: Comparison of clusterings of frequent regions and of individual cavities from serine protease structures. A) Clustering of frequent regions. B) Clustering of cavities from individual conformational samples. In both trees, topology is calculated based on volumetric distance. Coloring, which is independent of clustering topology, indicates the ligand binding preference of the protein.

volume of binding cavities in conformational samples of the entire dataset in Figure 6. The volume of trypsin cavities ranged from 248 $Å^3$ to 692 $Å^3$, the volume of chymotrypsin cavities ranged from 276 $Å^3$ to 568 $Å^3$, and elastase cavities ranged from 126 $Å^3$ to 552 $Å^3$, despite the general principle that chymotrypsin $S1$ cavities are larger to accommodate aromatic sidechains, and elastase cavities are smaller to accommodate amino acids like alanine or valine. The binding cavities of the enolase superfamily also exhibited substantial variation: Conformational samples of enolase cavities ranged from 90 $Å^3$ to 507 $Å^3$, mandelate racemases ranged from 225 $Å^3$ to 673 $Å^3$, and cavities sampled from muconate lactonizing enzyme were between 89 $Å^3$ and 343 $Å^3$. The substantial variation that can be observed between cavities illustrates the fundamental difficulty of accurately comparing binding site geometry in the presence of flexibility. Even though the variations are very small from the perspective of whole proteins, their effect on binding cavity geometry is substantial.

Using VASP-S [9], we trained a statistical model of structural variation between trypsin binding cavities and a second model of variation between enolase binding cavities. Using these two models, we evaluated the statistical significance of volumetric differences between pairs of trypsin cavities and between pairs of enolase cavities. Over 65 percent of CSG differences were incorrectly classified by VASP-S as being so large as to be consistent with different binding preferences.

These data indicate that conformational flexibility creates substantial variations between different conformational samples of the same protein. For this reason, it is apparent that similarities and variations in individual conformational samples can mislead comparisons and that techniques like FAVA, which compensate for conformational variation, are essential for cavity comparison.

3.2 UPGMA Clustering on frequent regions

A UPGMA clustering of frequent regions from the serine protease superfamily is plotted in Figure 7a. Trypsins were clustered tightly together, and elastases were correctly separated from trypsins, but Atlantic salmon elastase was placed as an outlier because it has zero volume. Chymotrypsin was separated, correctly, from both trypsins and elastases. In comparison, figure 7b illustrates one example of a UPGMA clustering generated from randomly selected conformational samples of each protein. While trypsins were correctly clustered together, salmon elastase (pdb: 1elt) was incorrectly classified as being more similar to the trypsins than it was to porcine elastase (pdb: 1b0e), and porcine elastase was incorrectly clustered with chymotrypsin.

Figure 8a illustrates a UPGMA clustering of frequent regions derived from enolase binding cavities.

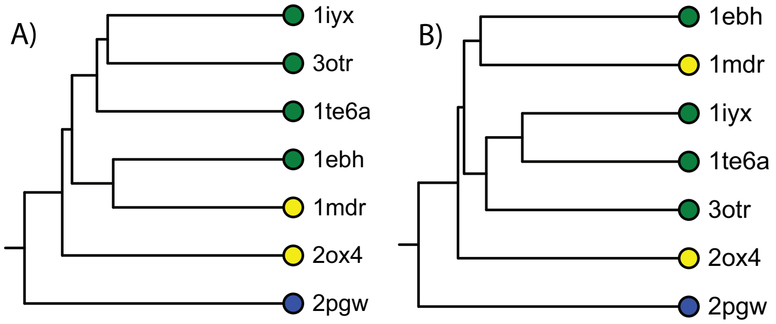

Figure 8: Comparison of clusterings of frequent regions and of individual cavities from enolase structures. A) Clustering of frequent regions. B) Clustering of cavities from individual conformational samples. In both trees, topology is calculated based on volumetric distance. Coloring, which is independent of clustering topology, indicates the ligand binding preference of the protein.

Frequent regions derived from muconate lactonizing enyzymes and enolases were separated correctly, as were frequent regions from mandelate racemase, with one exception: Mandelate racemase from Pseudomonas putida (pdb: 1mdr) was clustered with yeast enolase instead of with the other mandelate racemase from Zymomonas mobilis (pdb: 2ox4). A second UPGMA clustering, Fig 8b, was generated based on randomly selected conformational samples of each enolase superfamily structure. Individual conformational samples from the mandelate racemases in our dataset were distantly separated, indicating that their binding sites were quite different in these samples. Substantial differences were also apparent in baker's yeast enolase (pdb: 1ebh), which was separated from the enolase structures. Together, these results demonstrate that UPGMA clusterings of frequent regions represented similarities and differences in specificity. Overall, the flexible representation of binding cavities generated by FAVA exhibits fewer classification errors caused by conformational flexibility.

3.3 Influential amino acids

Differences in binding specificity can be caused by variations in the amino acids that surround the binding cavity, but these crucial variations can be obscured by sidechain and backbone motions. We hypothesized that amino acids that contribute to frequent differences between binding cavities, as opposed to those that create infrequent differences as a result of incidental motion, might also be responsible for differences in specificity. To evaluate this hypothesis, we computed the intersection volume between all residues r in all elastase structures (A), and all non-elastase serine protease cavities (B).

For example, the left of Figure 9 plots the degree of intersection between all conformational samples of all amino acids of porcine elastase (pdb: 1b0e) and binding cavities from conformational samples of salmon trypsin (pdb: 1bzx). Naturally, the vast majority of amino acids exhibited zero or very small intersection with the trypsin binding cavity because they are distant from the cavity when the two proteins are superposed. Two notable exceptions stand out: Samples of valine 216 and threonine 226 exhibited a median intersection volume of 45 Å^3 and 29 Å^3, respectively, with the conformational samples of the trypsin cavities. These amino acids are also known to block parts of the S1 pocket (inset, Figure 9), shortening it to accommodate small hydrophobic residues [50]. In this example, FAVA identified sterically influential elastase residues that have an established role in specificity.

We observed similar findings among the binding sites of the enolase superfamily (Figure 9, right). The aggregate intersection of conformational samples of amino acids from pseudomonas putida mandelate racemase (pdb: 1mdr) and binding cavities of saccharomyces cerevisiae enolase (pdb: 1ebh) offer a representative example: The vast majority of amino acids exhibited zero or very small intersection with the enolase binding cavity, but three exceptions are apparent: Samples of lysine 166, asparagine 197 and leucine 319 exhibited median intersection volumes of 14 Å^3, 13 Å^3, and 14 Å^3, respectively, with the conformational samples of the enolase cavities. Lysine 166 is believed to deprotonate (S)-mandelate substrates [31], asparagine 197 stabilizes the transition state [52], and leucine 319 assists in applying steric constraints in a hydrophobic pocket in the binding site [53]. Here, FAVA identified influential mandelate racemase residues that have an

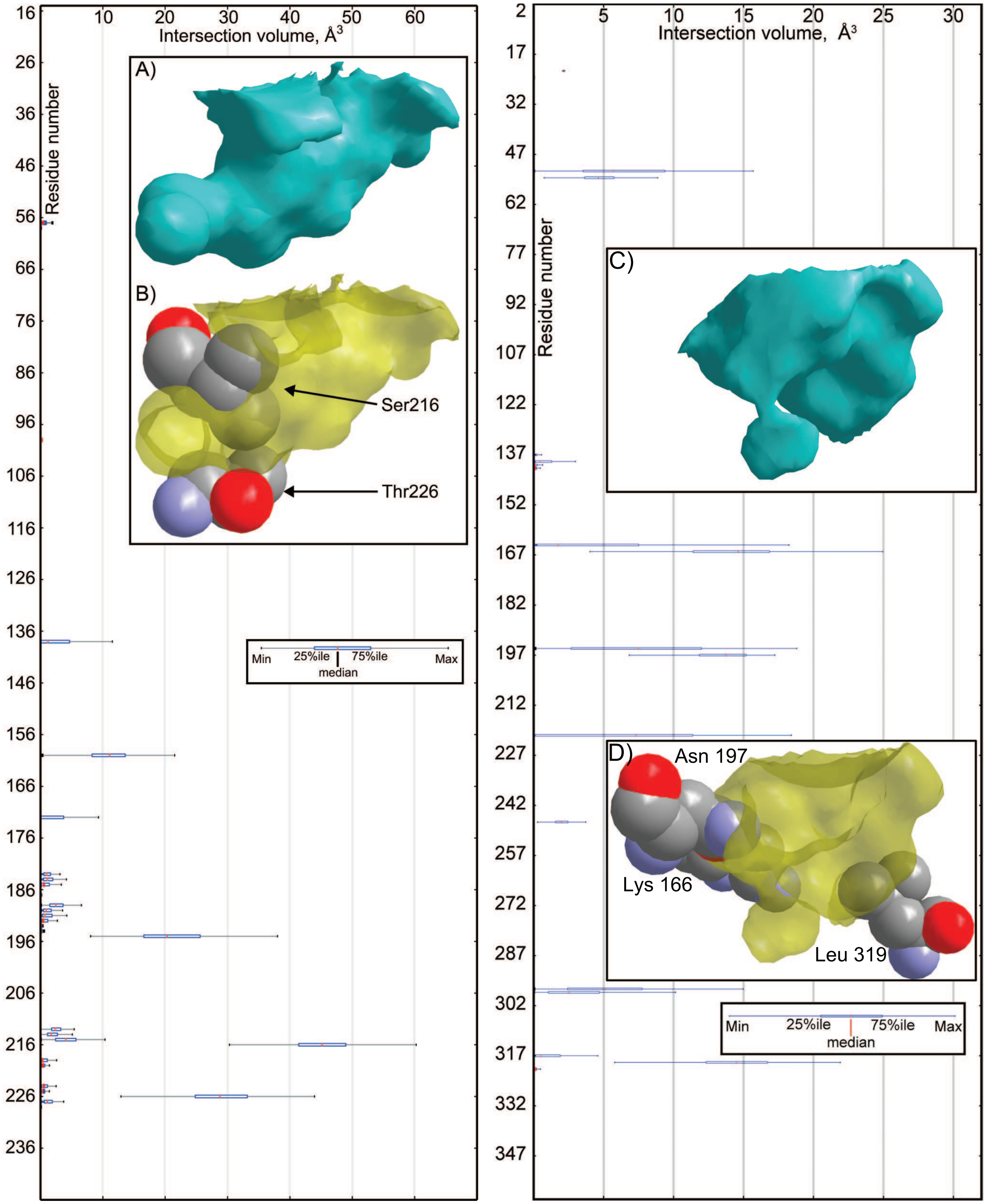

Figure 9: Left, a plot of the intersection volume of amino acids from conformational samples of porcine pancreatic elastase (pdb: 1b0e) with cavities from conformational samples of salmon trypsin (pdb: 1bzx). A) The trypsin cavity (teal). B) One snapshot of Val216 and Thr226 from 1b0e, relative to the cavity. Right, a plot of the intersection volume of amino acids from conformational samples of pseudomonas putida mandelate racemase (pdb: 1mdr) with cavities from conformational samples of saccharomyces cerevisiae enolase (pdb: 1ebh). C) The enolase cavity (teal). D) One snapshot of Lys166, Asp197, and Leu319 from 1mdr, relative to the cavity.

established role in function and specificity.

To calibrate the significance of these intersection volumes, we measured the volume of structural differences created by the motions of amino acids with binding cavities from the same protein. While side chain motion creates small variations, aggregate differences were considerably smaller. For example, among amino acids of porcine pancreatic elastase (pdb: 1b0e), the amino acid that most intersects the binding cavity of 1b0e is serine 195, the nucleophilic serine responsible for catalysis in all serine proteases [46]. It occupies an median of 5 Å^3 inside samples of binding cavities in 1b0e. This result suggests that the motions observed in the amino acids of the serine protease and enolase superfamilies create differences in binding site geometry that were far more significant than those that occur normally.

4 Discussion

We have presented FAVA, a method for incorporating conformational variations into the geometric comparison of protein binding cavities. Whereas existing representations describe large conformational changes in order to discover remote homologs in different conformations, FAVA is designed for the detailed comparison of local features, such as ligand binding cavities, in spite of interferance from small molecular motions, such as side chain flexibility. Whereas existing methods explicitly represent flexible regions of molecular structure, FAVA compiles volumetric representations of binding sites into frequent regions, which exclude infrequently occurring conformations and allow local conformational trends to emerge.

These conformational trends reflected differences in ligand binding preferences that could be obscured by conformational variation. A comparison of frequent regions generated from three subfamilies in each of the serine protease superfamily and the enolase superfamily revealed consistent variations in conformational trends that corresponded to differences in ligand binding preferences. If random conformational samples were compared instead of frequent regions, the resulting classification of binding sites could differ from actual binding preferences. Where conformational samples are selected arbitrarily, these results demonstrate the potential for errors in binding site comparison. Instead, using FAVA, we can mitigate these issues by classifying trends in a large sample of conformations.

Conformational trends also provided useful insights into the role of specific amino acids. We examined the range of conformations adopted by an amino acid over 100 nanoseconds and observed how frequently it would cause the binding site of a different protein to differ from its own. Amino acids that frequently cause differences in binding cavities almost always played a steric role in achieving binding specificity. While these influential residues occasionally move in ways that make different binding sites appear similar, their typical shape trends towards differences in binding sites, as might be expected based on different binding preferences. These results demonstrate that a detailed examination of many conformational samples can reveal individual influences on specificity and not simply trends in binding cavity shape.

FAVA has considerable potential for wider applications. First, it can be used with many sources of conformational samples, and as such it can provide insights into the patterns of conformational variations generated by many existing methods. As our capability to simulate molecular conformations continues to expand [48, 2], larger and more representative samples will enhance the accuracy of comaparisons made with FAVA. Second, in many cases, efforts to create proteins with engineered binding preferences already involve the simulation of protein structures. FAVA provides an analysis of the resulting simulation data that can contribute insights into the roles of individual amino acids. By using conformational samples to examine the detailed motions of ligand binding sites, sampled representations can offer important tools for the analysis of ligand binding specificity.

Acknowledgment

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

References

[1] Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., and Gerlt, J. A. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry 35*, 51 (1996), 16489–501.

[2] Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S., and Pande, V. S. Folding@ home: Lessons from eight years of volunteer distributed computing. In *Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on* (2009), IEEE, pp. 1–8.

[3] Berendsen, H., Postma, J., van Gunsteren, W., and Hermans, J. Intermolecular forces. *Pullman, B., Ed.; Reidel Publishing Company: Dordrecht* (1981), 331–342.

[4] Berglund, G., Smalas, A., Outzen, H., and Willassen, N. Purification and characterization of pancreatic elastase from North Atlantic salmon (Salmo salar). *Mol Mar Biol Biotechnol 7*, 2 (1998), 105–14.

[5] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res 28*, 1 (Jan. 2000), 235–42.

[6] Binkowski, T. A., Freeman, P., and Liang, J. pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Res 32, web server issue* (2004), W555–8.

[7] Birzele, F., Gewehr, J. E., Csaba, G., and Zimmer, R. Vorolign—fast structural alignment using voronoi contacts. *Bioinformatics 23*, 2 (2007), e205–e211.

[8] Bryant, D. H., Moll, M., Finn, P. W., and Kavraki, L. E. Combinatorial clustering of residue position subsets predicts inhibitor affinity across the human kinome. *PLoS computational biology 9*, 6 (2013), e1003087.

[9] Chen, B., and Bandyopadhyay, S. VASP-S: A Volumetric Analysis and Statistical Model for Predicting Steric Influences on Protein-Ligand Binding Specificity. In *Proceedings of 2011 IEEE International Conference on Bioinformatics and Biomedicine* (2011), pp. 22–9.

[10] Chen, B. Y., Fofanov, V. Y., Bryant, D. H., Dodson, B. D., Kristensen, D. M., Lisewski, A. M., Kimmel, M., Lichtarge, O., and Kavraki, L. E. The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs. *J Comp Biol 14*, 6 (2007), 791–816.

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

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

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

[14] DeLano, W. L. The PyMOL Molecular Graphics System, 2002.

[15] Dundas, J., Adamian, L., and Liang, J. 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 (2011), 713–729.

[16] Felsenstein, J. Phylip - phylogeny inference package (version 3.2). 164–166.

[17] Godshall, B. G., and Chen, B. Y. Improving accuracy in binding site comparison with homology modeling. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on* (2012), IEEE, pp. 662–669.

[18] Godshall, B. G., Tang, Y., Yang, W., and Chen, B. Y. An aggregate analysis of many predicted structures to reduce errors in protein structure comparison caused by conformational flexibility. *BMC structural biology 13*, Suppl 1 (2013), S10.

[19] Gráf, L., Jancsó, a., Szilágyi, L., Hegyi, G., Pintér, K., Náray-Szabó, G., Hepp, J., Medzihradszky, K., and Rutter, W. J. Electrostatic complementarity within the substrate-binding pocket of trypsin. *Proc Natl Acad Sci U S A 85*, 14 (July 1988), 4961–5.

[20] Gunasekaran, K., and Nussinov, R. How Different are Structurally Flexible and Rigid Binding Sites ? Sequence and Structural Features Discriminating Proteins that Do and Do not Undergo Conformational Change upon Ligand Binding. *J Mol Biol 365* (2007), 257–273.

[21] Guo, Z., and Chen, B. Y. Variational bayesian clustering on protein cavity conformations for detecting influential amino acids. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (2014), ACM, pp. 703–710.

[22] Guo, Z., Kuhlengel, T., Stinson, S., Blumenthal, S., Chen, B. Y., and Bandyopadhyay, S. A flexible volumetric comparison of protein cavities can reveal patterns in ligand binding specificity. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (2014), ACM, pp. 445–454.

[23] Hess, B. P-LINCS: a parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation 4*, 1 (Jan. 2008), 116–122.

[24] Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation 4*, 3 (Mar. 2008), 435–447.

[25] Holm, L., and Sander, C. The fssp database of structurally aligned protein fold families. *Nucleic acids research 22*, 17 (1994), 3600.

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

[27] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics 79*, 2 (1983), 926–935.

[28] Junier, T., and Zdobnov, E. M. The newick utilities: high-throughput phylogenetic tree processing in the unix shell. *Bioinformatics 26*, 13 (2010), 1669–1670.

[29] Konc, J., and Janežič, D. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics 26*, 9 (2010), 1160–1168.

[30] Kühnel, K., and Luisi, B. F. Crystal structure of the Escherichia coli RNA degradosome component enolase. *J Mol Biol 313*, 3 (Oct. 2001), 583–92.

[31] Landro, J. A., Gerlt, J. A., Kozarich, J. W., Koo, C. W., Shah, V. J., Kenyon, G. L., Neidhart, D. J., Fujita, S., and Petsko, G. A. The role of lysine 166 in the mechanism of mandelate racemase from pseudomonas putida: Mechanistic and crystallographic evidence for stereospecific alkylation by (r)-. alpha.-phenylglycidate. *Biochemistry 33*, 3 (1994), 635–643.

[32] Menke, M., Berger, B., and Cowen, L. Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology 4*, 1 (2008), e10.

[33] Morihara, K., and Tsuzuki, H. Comparison of the specificities of various serine proteinases from microorganisms. *Arch Biochem Biophys 129*, 2 (1969), 620–634.

[34] Mosca, R., and Schneider, T. R. Rapido: a web server for the alignment of protein structures in the presence of conformational changes. *Nucleic acids research 36*, suppl 2 (2008), W42–W46.

[35] Murzin, a. G., Brenner, S. E., Hubbard, T., and Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247*, 4 (Apr. 1995), 536–40.

[36] Nose, S., and Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics 50*, 5 (1983), 1055–1076.

[37] Nussinov, R., and Wolfson, H. J. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A 88*, 23 (Dec. 1991), 10495–9.

[38] Orengo, C. a., Michie, a. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. CATH–a hierarchic classification of protein domain structures. *Structure 5*, 8 (Aug. 1997), 1093–108.

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

[40] Parrinello, M., and Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics 52* (1981), 7182.

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

[42] Rakus, J. F., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Hubbard, B. K., Delli, J. D., Babbitt, P. C., Almo, S. C., and Gerlt, J. A. Evolution of enzymatic activities in the enolase superfamily: L-rhamnonate dehydratase. *Biochemistry 47*, 38 (Sept. 2008), 9944–54.

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

[44] Schaer, J., and Stone, M. Face traverses and a volume algorithm for polyhedra. *Lect Notes Comput Sc 555/1991* (1991), 290–297.

[45] Schafer, S. L., Barrett, W. C., Kallarakal, A. T., Mitra, B., Kozarich, J. W., Gerlt, J. A., Clifton, J. G., Petsko, G. A., and Kenyon, G. L. Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the D270N mutant. *Biochemistry 35*, 18 (May 1996), 5662–9.

[46] Schechter, I., and Berger, A. On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications 27*, 2 (1967), 157–162.

[47] Shatsky, M., Nussinov, R., and Wolfson, H. J. FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. *J Comput Biol 11*, 1 (Jan. 2004), 83–106.

[48] Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., et al. Atomic-level characterization of the structural dynamics of proteins. *Science 330*, 6002 (2010), 341–346.

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

[50] Shotton, D., and Watson, H. Three-dimensional structure of tosyl-elastase. *Nature 225*, 5235 (1970), 811–816.

[51] Sneath, P. H., and Sokal, R. R. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

[52] St. Maurice, M., and Bearne, S. L. Reaction intermediate analogues for mandelate racemase: interaction between asn 197 and the α-hydroxyl of the substrate promotes catalysis. *Biochemistry 39*, 44 (2000), 13324–13335.

[53] St. Maurice, M., and Bearne, S. L. Hydrophobic nature of the active site of mandelate racemase. *Biochemistry 43*, 9 (2004), 2524–2532.

[54] Stark, A., Sunyaev, S., and Russell, R. B. A Model for Statistical Significance of Local Similarities in Structure. *J Mol Biol 326* (2003), 1307–1316.

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

[56] Vesterstrøm, J., and Taylor, W. R. Flexible secondary structure based protein structure comparison applied to the detection of circular permutation. *Journal of Computational Biology 13*, 1 (2006), 43–63.

[57] Yang, A.-S., and Honig, B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol 301*, 3 (Aug. 2000), 665–78.

[58] Ye, Y., and Godzik, A. Multiple flexible structure alignment using partial order graphs. *Bioinformatics 21*, 10 (May 2005), 2362–9.

[59] Zhang, Y., and Skolnick, J. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research 33*, 7 (2005), 2302–2309.