OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

VASP-E: Specificity Annotation with a Volumetric Analysis of Electrostatic Isopotentials

Brian Y. Chen*

Department of Computer Science and Engineering, P.C. Rossin College of Engineering and Applied Sciences, Lehigh University, Bethlehem, Pennsylvania, United States of America

Abstract

Algorithms for comparing protein structure are frequently used for function annotation. By searching for subtle similarities among very different proteins, these algorithms can identify remote homologs with similar biological functions. In contrast, few comparison algorithms focus on specificity annotation, where the identification of subtle differences among very similar proteins can assist in finding small structural variations that create differences in binding specificity. Few specificity annotation methods consider electrostatic fields, which play a critical role in molecular recognition. To fill this gap, this paper describes VASP-E (Volumetric Analysis of Surface Properties with Electrostatics), a novel volumetric comparison tool based on the electrostatic comparison of protein-ligand and protein-protein binding sites. VASP-E exploits the central observation that three dimensional solids can be used to fully represent and compare both electrostatic isopotentials and molecular surfaces. With this integrated representation, VASP-E is able to dissect the electrostatic environments of protein-ligand and protein-protein binding interfaces, identifying individual amino acids that have an electrostatic influence on binding specificity. VASP-E was used to examine a nonredundant subset of the serine and cysteine proteases as well as the barnase-barstar and Rap1a-raf complexes. Based on amino acids established by various experimental studies to have an electrostatic influence on binding specificity, VASP-E identified electrostatically influential amino acids with 100% precision and 83.3% recall. We also show that VASP-E can accurately classify closely related ligand binding cavities into groups with different binding preferences. These results suggest that VASP-E should prove a useful tool for the characterization of specific binding and the engineering of binding preferences in proteins.

Citation: Chen BY (2014) VASP-E: Specificity Annotation with a Volumetric Analysis of Electrostatic Isopotentials. PLoS Comput Biol 10(8): e1003792. doi:10.1371/journal.pcbi.1003792

Editor: Dennis R. Livesay, UNC Charlotte, United States of America

Received February 21, 2014; **Accepted** June 17, 2014; **Published** August 28, 2014

Funding: This work was supported initially by National Institutes of Health grants GM030518 and GM094597 to Barry Honig and later by National Science Foundation Grant 1320137 to BYC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* Email: chen@cse.lehigh.edu

This is a *PLOS Computational Biology* Methods article.

Introduction

Software for comparing protein structures is widely used to make inferences about protein function. These methods assist in function annotation by revealing proteins that perform similar biological functions despite vast evolutionary differences. Many methods focus on the discovery of subtle structural similarities among very different molecules [1,2] using the superposition of catalytic residues [3–5] or the comparison of binding cavities [6–9]. By aligning polypeptide backbones [10–14], distance matrices [15] or geometric graphs [16–18], related methods can reveal similarities in tertiary structure that are not evident from sequences alone. Most approaches use atom coordinates or molecular surfaces [19–22] as digital representations of protein geometry. Other characteristics, such as evolutionary significance [23–25], hydrophobicity [26] and electrostatic potential [5,23,27] are attached to this geometric representation as labels. Comparisons of these data often generate a score, such as the root mean squared distance (RMSD), that summarizes structural, biological, and chemical similarities among two or more structures. Proteins with very different sequences sometimes exhibit unusually similar RMSDs, revealing shared origins in antiquity [28–30].

An emerging second type of comparison algorithm is designed to find subtle differences among very similar proteins. These methods seek to annotate protein specificity by proposing structural causes for different binding preferences among proteins that perform the same function [31–36]. For example, specificity annotation software might identify a cleft inside the ligand binding cavity of one protein that does not exist in a close homolog. That cleft might accommodate substrates that the homolog cannot bind. To find structural features like these, RMSD, and other scores for function annotation, are inappropriate because they obscure meaningful individual variations, like the cleft, behind summary scores. Instead, volumetric representations [37], which describe protein structure based on spatial regions occupied by the atoms of a protein, can be used to identify individual structural differences that may alter the binding preferences of ligand binding cavities [31–34]. To date, few comparisons focused on finding subtle electrostatic differences among closely related proteins have been reported, even though electrostatic fields are widely used to infer protein function [38–47] and specificity [48–53]. To fill this gap, this paper proposes a novel volumetric representation and comparison algorithm for finding electrostatic influences on binding specificity.

The problem we are specifically addressing is the case where several closely related proteins have already been structurally

Author Summary

Proteins, the ubiquitous worker molecules of the cell, are a diverse class of molecules that perform very specific tasks. Understanding how proteins achieve specificity is a critical step towards understanding biological systems and a key prerequisite for rationally engineering new proteins. To examine electrostatic influences on specificity in proteins, this paper presents VASP-E, a software tool that generates solid representations of the electrostatic potential fields that surround proteins. VASP-E compares solids with constructive solid geometry, a class of techniques developed first for modeling complex machine parts. We observed that solid representations could quantify the degree of charge complementarity in protein-protein interactions and identify key residues that strengthen or weaken them. VASP-E correctly identified amino acids with established experimental influences on protein-protein binding specificity. We also observed that solid representations of electrostatic fields could identify electrostatic conservations and variations that relate to similarities and differences in binding specificity between proteins and small molecules.

aligned and we seek to identify spatially conserved and varying regions in their potential fields that might cause differences in binding specificity. Conserved regions, where the fields have similar potentials, might stabilize a molecular fragment attracted by all proteins (Fig. 1g), while differences in specificity could arise from regions where the fields vary (Fig. 1h,i). Software for identifying conservation and variation in charged regions can thus suggest how such regions may play a role in molecular recognition, and how they might be changed to achieve different binding preferences. Our approach identifies regions like these by representing electrostatic isopotentials with volumetric solids generated by the new program VASP-E (Volumetric Analysis of Surface Properties with Electrostatics). VASP-E computes conserved and varying regions using techniques from Constructive Solid Geometry (CSG) (Fig. 1). Developed originally for computer aided design [54] and computer graphics [55], CSG enables unions, intersections, and differences of volumetric representations to be calculated as if they are three dimensional solids. When used to analyze fields, CSG intersections can approximate regions that are common to isopotentials from several aligned proteins, thereby identifying regions with conserved potentials. CSG differences identify regions inside the isopotential of one protein but not inside that of another, isolating a region where potentials vary. Together, CSG operations provide a novel mechanistic approach to the analysis of electrostatic fields because the approximation of conserved and varying fields is not possible with existing structure comparison methods.

The solid representations employed by VASP-E differ in kind from existing electrostatic analyses. While VASP-E deconstructs the electrostatic field to identify conserved and varying electrostatic phenomena, existing methods summarize and quantify the field with comparison scores [56,57] and biophysical energies [58–61]. These numerical values cannot point to specific regions in the field with electrostatic similarities or differences, and so they cannot suggest how a protein might be altered to engineer different binding preferences. A second fundamental difference is that solid representations have the additional capability to represent the region inside molecular surfaces. Using CSG, we can therefore integrate both types of data to focus on electrostatic fields within binding sites. For example, the CSG difference of an isopotential minus the molecular surface at a binding site represents a three dimensional charged region in the solvent that can be occupied by potential binding partners (e.g. Fig. 1c,d). In contrast, representations used in function annotation methods generally represent electrostatic fields at or near the molecular surface only. Sampling a three dimensional field along this curving two dimensional surface cannot describe the electrostatic field as it extends outwards from the protein and influences other molecules. Third, while existing methods characterize fields at all potentials, solid representations describe fields at selected isopotential thresholds only. This feature enables comparisons to focus on ranges of potential that are especially relevant to binding. It can also be used to measure electrostatic complementarity between binding partners, as we will demonstrate later, by identifying interface regions where oppositely charged isopotentials overlap. To our knowledge, VASP-E is the first application of CSG to the volumetric comparison of electrostatic isopotentials, although tree-based methods that summarize topological differences in electrostatic isopotentials [57] have also been developed.

This paper explores two applications of VASP-E as it might be applied in support of research in structural biology. One objective in many investigations is to discover electrostatic influences on protein-ligand or protein-protein binding specificity. Given the long range nature of electrostatic interactions, many amino acids could potentially be influential, and it could be impractical to create all possible mutants and determine their binding preferences. Here, a first application of VASP-E is to suggest amino acids that create differences between the electrostatic fields of two ligand binding cavities or to suggest amino acids that enhance or diminish electrostatic complementarity between two interacting proteins. Because amino acids are suggested in tandem with a hypothetical electrostatic influence on binding, VASP-E provides reasons to produce and test certain mutants first, where no reason might have existed before. The second application of VASP-E examined in this paper is the classification of protein-ligand binding cavities based on their electrostatic fields. This application can support efforts to discover patterns of electrostatic similarities or differences among related binding sites. In studies seeking to identify a possible ligand, electrostatic classification can reveal similarities to other proteins that may have known binding partners. Together, these applications of VASP-E represent two of many capabilities that become possible by combining CSG and volumetric representations of electrostatic isopotentials. We validate these capabilities in the results section against established experimental observations.

Methods

2.0.1 Method summary

The underlying observation exploited by VASP-E is that geometric comparisons of electrostatic potential fields can focus on biologically relevant regions and specific potential ranges by using CSG. Constraining the comparison of potential fields in this manner ensures that comparisons reflect aspects of electrostatic fields that influence binding, rather than spurious variations that occur by random chance or outside of binding sites. To achieve this kind of focus, comparisons always begin with a multiple structure alignment of whole proteins [10–18,62], where ligand binding cavities or protein-protein interfaces are defined on each structure using cavity detection algorithms [63–67] or manual design.

Structures aligned in this manner are then used to generate solid representations of electrostatic isopotentials and protein structure. To represent electrostatic isopotentials, we first solve the potential

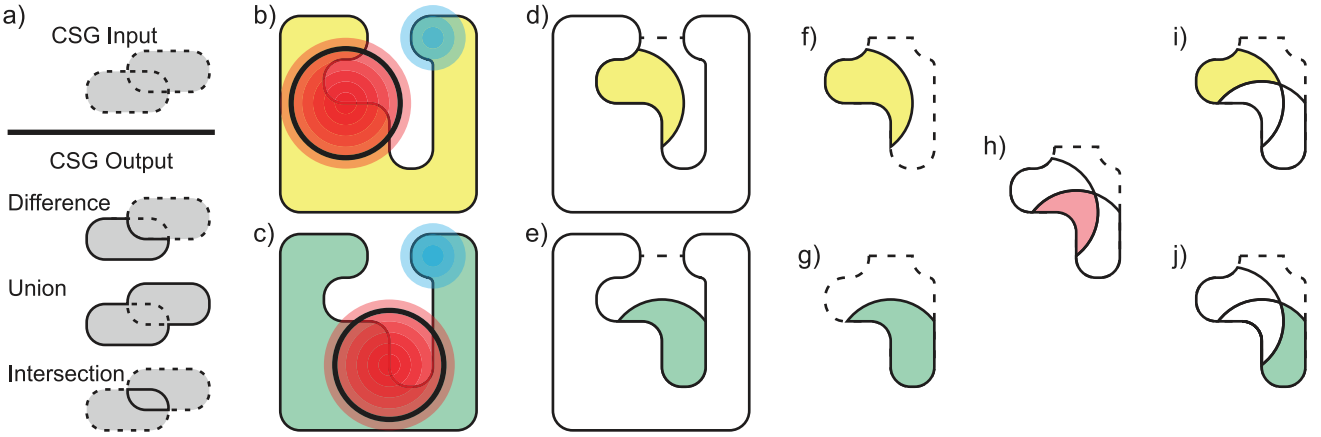

Figure 1. CSG analysis of electrostatic isopotentials in ligand binding cavities. a) A demonstration of CSG operations, illustrating the borders of input (dotted) and output (solid) regions in grey (grey everywhere). b,c) Shapes representing the regions occupied by protein **X** (yellow) and **Y** (green), their molecular surfaces (thin black lines), and their electrostatic potential fields (red and blue gradients). Regions with increasingly negative potential are shown in darker red, and regions with increasingly positive potential are shown in darker blue. An isopotential selected by a user is shown with a heavy black line. d,e) The CSG differences **x** and **y** between the region within the user-selected isopotential and the molecular surface of **X** and **Y** is shown in yellow and green. These volumes represent solvent accessible cavity regions with electrostatic potential at least as negative as that selected by the user. The external boundary of the ligand binding cavities of **X** and **Y** is shown with a dotted line. f,g) **x** and **y** are shown in yellow and green, with a black boundary. The ligand binding cavities they occupy are shown with a dotted boundary. h) The CSG intersection (red) of **x** and **y** (black outlines), when **X** and **Y** are aligned, represents a solvent accessible cavity region where electrostatic potential in both proteins is at least as negative as that selected by the user. i,j) The CSG differences of **x** - **y** (yellow) and **y** - **x** (green), indicating regions of electrostatic potential in the ligand binding cavity of one protein and not the other.
doi:10.1371/journal.pcbi.1003792.g001

field of a given structure using DelPhi [68]. Using the field, isopotential surfaces are approximated using Marching Cubes [69], an algorithm first applied to visualize electrostatic isopotentials in GRASP [70]. This method is paraphrased below. Solids representing molecular surfaces are generated using the Trollbase library [12], which implements the classical rolling-probe method [71].

The resulting solids, regardless of their origin, are basic inputs for CSG operations, which we described earlier [37]. Below, we use the symbols $\cap$, $\cup$ and $-$ to denote intersection, union, and difference operations, which are the basic CSG operations used in this work. VASP-E uses CSG to integrate solid representations of electrostatic isopotentials and molecular surfaces to create solid representations of the electrostatic field within ligand binding cavities (*cavity fields*) and protein-protein interfaces (*interface fields*). These procedures are detailed below.

Cavity fields and interface fields are the constrained representations used by VASP-E to focus the comparison of electrostatic fields on biologically significant regions. To quantify similarities, we compute the CSG intersection of two regions and then evaluate the volume of the resulting intersection region. To quantify differences, we measure the volume of the CSG difference. Large volumes of intersection imply similar fields while large differences are characteristic of fields that vary. To estimate the volume $v(X)$ of any region X, including outputs from CSG operations, we use the Surveyor's Formula [72], which we described earlier [37].

Further CSG operations permit deconstructive comparisons of cavity and interface fields that identify similarities in some regions and differences in other regions within the fields they describe. While many applications this kind are possible with VASP-E, we describe two below: First, we can use VASP-E to trace differences in electrostatic fields to individual amino acids that contribute to these differences, thereby predicting residues that influence binding specificity. Second, we can integrate multiple electrostatic similarity measurements between a family of cavity fields to reveal patterns of ligand binding specificity.

2.1 Solid representations of electrostatic isopotentials with marching cubes

As input, marching cubes begins with a molecular structure from the Protein Data Bank (PDB) [73], its electrostatic potential field E, a desired isopotential threshold k, and the user's choice of representing the region with potential greater than or less than k. The overall procedure (Fig. 2) approximates the solid region on one side of the isopotential at k, which we refer to as a *solid isopotential* (Fig. 2a). In this work, when generating electrostatic isopotentials at k kT/e, we always represent the region with potential greater than k when k is positive, and the region with potential less than k, when k is negative. Regions on the other sides of these potentials are infinite in volume, and thus their comparison is not well defined. Below, we use a negative value for k and represent the region on the lower-potential side of k, as an example.

First, we protonate the PDB structure using the *reduce* component of MolProbity [74]. The resulting structure is passed to DelPhi [68], which computes numerical solutions to the nonlinear Poisson-Boltzmann equation, yielding an approximation of E at every point within a bounding box surrounding the protein. Using E, Marching Cubes outputs a polyhedral approximation of the isopotential surface at k kT/e, which we interpret as the exterior boundary of a three dimensional solid.

Marching Cubes begins by establishing a regular lattice of cubes around the protein, whose borders fall within the bounding box (Fig. 2b). The lattice as a whole can be interpreted as a collection of *lattice points* at the corners of each cube, *lattice edges* connecting adjacent corners, *lattice faces* between cubes, or as simply a collection of *lattice cubes*. The *resolution* of the lattice, defined by the length of a lattice edge, is specified by the user and can be changed to accommodate structures of different sizes in system memory.

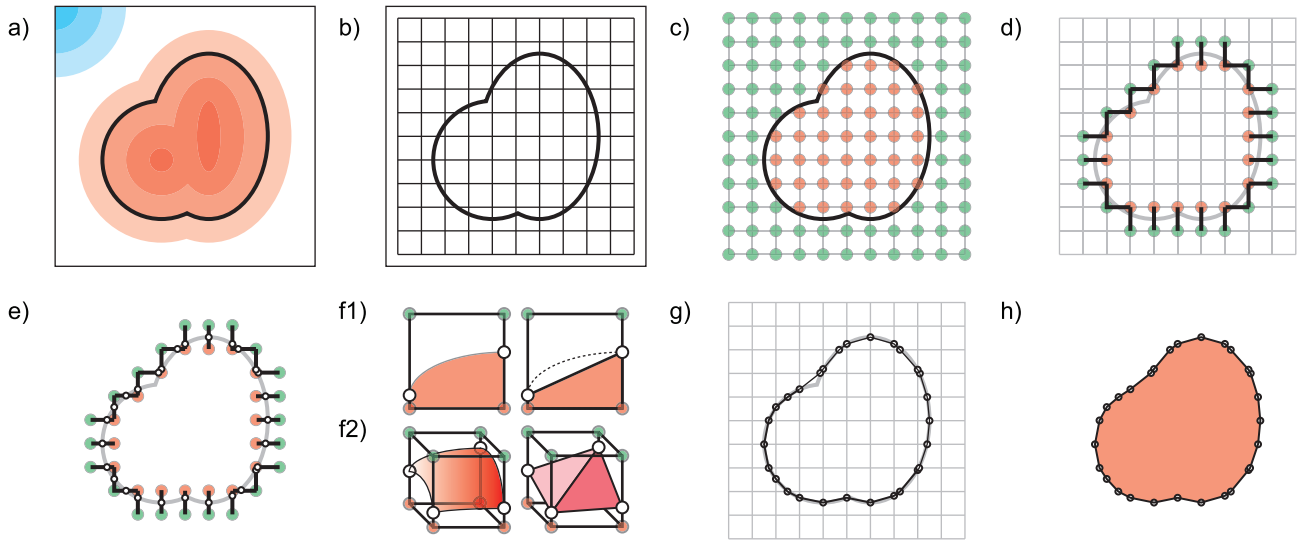

Figure 2. Generating a solid representation of an electrostatic isopotential using marching cubes. a) The input electrostatic field, illustrated as a gradient of red (negative potential) and blue (positive potential) regions. The solid region to be approximated is within the heavy black line. b) Axis aligned cubic lattice surrounding solid isopotential (black grid). c) Lattice points (circles) evaluated as being inside (red) or outside (green) the isopotential. d) Selected edges, found between interior and exterior lattice points (short black lines), intersect the electrostatic isopotential (grey curved line). e) Intersection points along each selected edge (small white circles). f1) A two dimensional illustration of the solid isopotential passing through a lattice square (red, left), with interior lattice points shown with red circles, and exterior lattice points shown with green circles. An approximation of the solid isopotential using a straight line is shown on the right. f2) A three dimensional illustration of the surface of a solid isopotential (red gradient, left) inside a lattice cube. Lattice points inside the solid isopotential are shown as red circles, lattice points outside are shown in green. An approximation of the solid isopotential triangles connecting intersection points (white circles) is shown on the right. g) Together, the triangles in all cubes (black lines) form the boundary surface approximating the solid isopotential (h).
doi:10.1371/journal.pcbi.1003792.g002

Once the lattice is initialized, we evaluate the potential $U_E(r)$ of the field E at every lattice point r. If $U_E(r) \leq k$, we mark r as being *inside* the isopotential. Otherwise, we mark r as being *outside* (Fig. 2c). The evaluation of $U_E(r)$ is made possible using the Trollbase library [12], which evaluates the field to determine the potential at r.

Next, we select every lattice edge that connects an inside lattice point to one outside. Since isopotentials are topologically closed surfaces, the selected edge must intersect the desired isopotential (Fig. 2d). On each selected edge, we estimate the *intersection point* p between the segment and the isopotential using linear interpolation between the electrostatic potentials at the endpoints (Fig. 2e).

Finally, we consider every lattice cube joined to at least one lattice edge with an intersection point. On the cube, the intersection points collectively approximate the places where the isopotential passes through the cube. In two dimensions, this can be drawn as a shape passing through a square (Fig. 2f1), and approximated with a line through the cube. In three dimensions, when the isopotential passes through the cube, it's boundary is a surface that intersects the segments at the intersection points calculated earlier. This surface can be approximated inside each cube with triangles connecting triplets of intersection points. We use a lookup table to specify each triangle layout based on the $2^8 = 256$ possible combinations of selected edges in a cube (Fig. 2f2). Assembling the triangles into a single surface generates the exterior boundary of the solid isopotential (Fig. 2g,h).

2.2 Generating and comparing cavity fields

A cavity field is a solid representation of the region inside a ligand binding cavity that is also inside a solid isopotential. To generate a cavity field, we require the solid isopotential and a solid representation of the ligand binding cavity (Fig. 3a,b). We generate the solid isopotential using the method above and represent the ligand binding cavity using VASP and a volumetric approach based on SCREEN [65], described earlier [37]. Computing a CSG intersection of these two regions generates the cavity field (Fig. 3c,d).

We compare cavity fields to detect local electrostatic differences that might affect specificity. Our approach follows the assumption that the user has selected solid isopotentials at a threshold that is relevant for ligand binding. For example, if a negative potential is influential for the selection of positively charged substrates, comparing regions of negative potential in several cavities could reveal electrostatic causes for different binding preferences. We discuss the selection of these potentials in Supplemental Text S1.

Our comparison begins by structurally aligning two proteins, A and B, and generating their cavities, A_P and B_P. Because A_P and B_P are regions that are outside the molecular surface, we say that they are *solvent accessible regions*. Using A_P and B_P, we generate their cavity fields, A_C and B_C at the same side of the electrostatic potential k. Next, we generate the intersection $I = A_P \cap B_P$. I is the region that is solvent accessible in both cavities (Fig. 3e). By comparing electrostatic fields inside I, we are guaranteed that our comparison is not influenced by steric differences. For this reason, we next compute the intersection $I_A = I \cap A_C$ and $I_B = I \cap B_C$ (Fig. 3f,g). I_A and I_B are regions within the solid isopotentials of A_C and B_C that are solvent accessible in both cavities.

Computing I_A and I_B permits several useful comparisons. First, the intersection $I_A \cap I_B$ (Fig. 3h), is solvent accessible in both cavities and also inside both solid isopotentials. This region of structural and electrostatic similarity might stabilize molecular fragments that are common to substrates of both proteins. Second, the difference regions $D_{A\,B} = I_A - I_B$ and $D_{B\,A} = I_B - I_A$ (Fig. 3i,j)

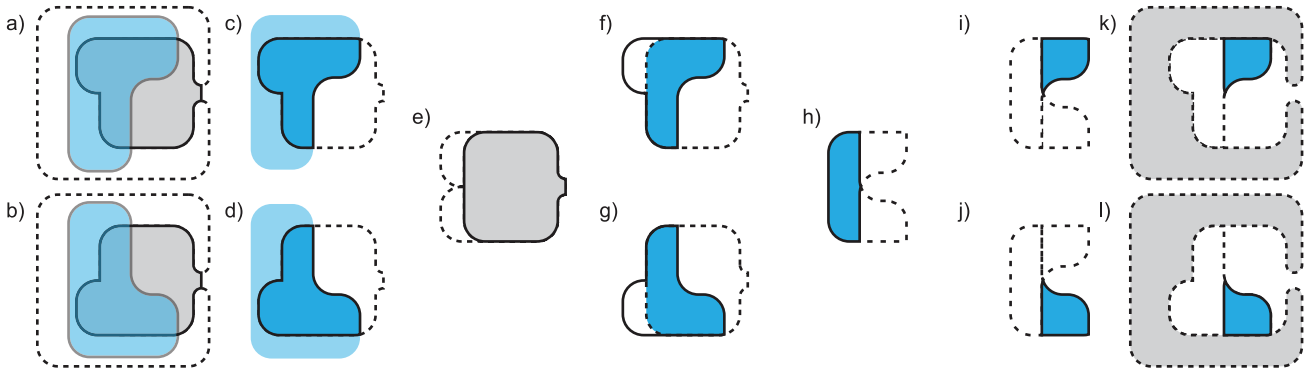

Figure 3. Cavity field generation and comparison. a,b) The molecular surfaces of protein A and protein B are shown in dotted outlines. Ligand binding cavities are shown in grey with solid outlines. A solid isopotential is shown in transparent blue. c,d) Cavity fields A_C and B_C are shown in opaque blue, with the solid isopotential (transparent blue), and the binding cavity (dotted outline). e) The intersection of the two binding cavities, I (grey) is a solvent accessible region in both cavities. f,g) The intersections I_A and I_B are shown in blue, I is shown with a dotted outline, and cavity fields A_C and B_C (solid outline). h) I_A and I_B (dotted outline) and their intersection (blue). i,j) I_A and I_B shown with a dotted outline, with differences shown in blue. k,l) The same differences relative to the molecular surface.
doi:10.1371/journal.pcbi.1003792.g003

are solvent accessible in both cavities but different in electrostatic character, because they lie inside the solid isopotential of one cavity and not the other. Molecular fragments in this region may thus be accommodated by one protein and electrostatically destabilized in the other.

We quantify differences by measuring the volume of $D_{A\,B}$ and $D_{B\,A}$. If $v(D_{A\,B})$ and $v(D_{B\,A})$ are small, we say that the cavity fields are similar. If one or both volumes are large, we say that A_C and B_C are electrostatically dissimilar, and that $D_{A\,B}$ (or $D_{B\,A}$) is evidence supporting the hypothesis that A_C could attract or stabilize a ligand that B_C cannot. This computation enables a systematic categorization of all electrostatic differences in the binding cavities of A and B.

2.3 Generating and comparing interface fields

An interface field is a solid representation of a region of electrostatic complementarity between two proteins A and B, in complex. Given a potential threshold k, we define a region of *electrostatic complementarity* to be the spatial region where the field of A, independent of B, has potential greater than k, and the field of B, independent of A, has potential less than $-k$. We refer to this region as an *interface field*. To generate an interface field, we require three inputs: A solid representation of the interface region, an electrostatically significant isopotential of A alone at k kT/e (A_k), and an electrostatically significant isopotential of B alone at $-k$ kT/e (B_{-k}). Because we use interface fields to analyze the specificity of interacting proteins, and because VASP-E is not designed to determine how two proteins interact, unbound structures are not used for the generation of interface fields.

To generate the interface region, we first identify amino acids at the interface (Fig. 4a). These are the amino acids of A that have an atom within 5 Å of B, and the amino acids of B that have an atom within the same distance of A. Next, we generate spheres with radius 5 Å, centered at every atom of these amino acids (Fig. 4b). Finally, we compute the interface region I with the CSG union of these spheres (Fig. 4c). In identifying amino acids that are part of the interface region, we do not include or exclude amino acids based the fraction of their surface exposed to the solvent, because some influential "hot spot" amino acids may have low solvent exposure [75].

Electrostatically significant isopotentials A_k and B_{-k} (Fig. 4d,e) are generated with the marching cubes method described above. Using A_k and B_{-k}, we compute $A_k \cap I$, the electrostatically significant region of the field of protein A within the interface, and $B_{-k} \cap I$, the oppositely charged electrostatically significant region of the field of protein B within the interface (Fig. 4f,g). The intersection of these two regions is the interface field, $IF_{A\,B}(k)$ (Fig. 4h).

Since the interface field represents electrostatic complementarity in a given complex, we can use interface fields to compare electrostatic complementarity in two complexes. For two complexes, C and D, and k, the user's threshold for electrostatic significance, we generate four interface fields: $IF_C(k)$, $IF_C(-k)$, $IF_D(k)$, and $IF_D(-k)$. Comparing the interface fields at k and $-k$ yields a more complete representation of electrostatic complementarity in both complexes.

We evaluate the difference d between two complexes using the following expression:

$$d(C,D) = (v(IF_C(k)) + v(IF_C(-k))) - (v(IF_D(k)) + v(IF_D(-k)))$$

Where $v(Y)$ denotes the volume of a given region Y. The two interface fields for each complex express the degree of complementarity on the positive and negative parts of the electrostatic potential spectrum. The interface fields for the same complex are summed, to represent the total degree of complementarity on that complex. The difference between the two sums expresses the difference in complementarity d between the two complexes, on both sides of the potential spectrum. Large absolute values of d indicate large differences in electrostatic complementarity between the two complexes, while values close to zero point to similar degrees of complementarity.

2.4 Identifying electrostatically influential amino acids

DelPhi [68] is able to solve the electrostatic field of a given protein structure while omitting the electrostatic contribution of a individual amino acid. This process, which we refer to as *nullification*, has the unique property of leaving the structure of the amino acid intact while eliminating its electrostatic contribution. Maintaining the structure of the protein is important in an electrostatic analysis because the nullified amino acid still displaces solvent, creating a region of low dielectric. That region can enhance the electrostatic potentials of amino acids that were not

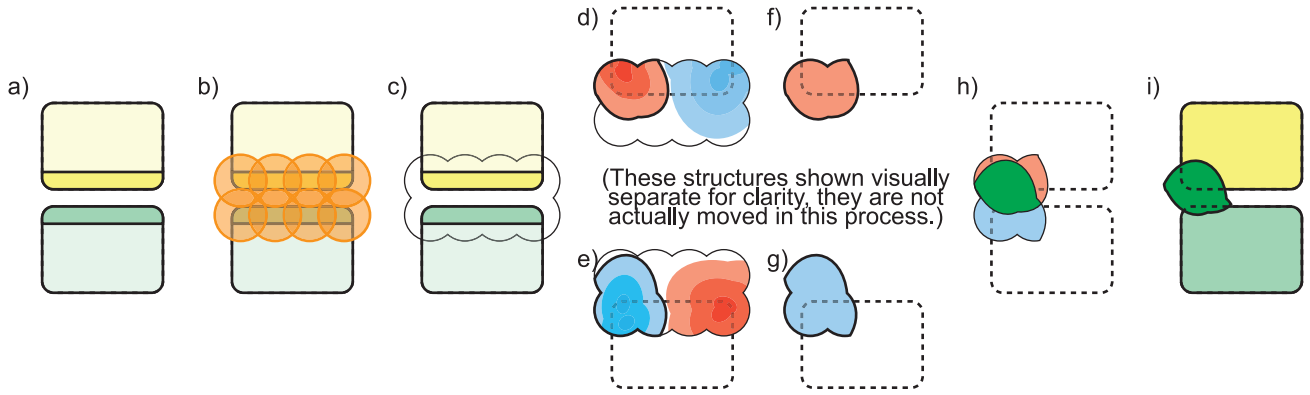

Figure 4. Generating an interface field. a) Two proteins in complex (rounded rectangles), and amino acids at the protein-protein interface (yellow, green). b) Spheres around every atom in the interfacial amino acids (orange). c) CSG union of interfacial spheres. d,e) Red and blue gradients representing the electrostatic potential field in the interfacial regions of protein A (d) and B (e). Black lines represent the user-selected isopotential at k in protein A and $-k$ in protein B. f,g) The electrostatically significant isopotentials from A and B, in red and blue. h) The CSG intersection (green) of the isopotentials from A and B. i) The interface field (transparent green).
doi:10.1371/journal.pcbi.1003792.g004

nullified because of an effect called *electrostatic focusing*. Electrostatic focusing is known to play a considerable role in function and specificity [48,76,77]. Below, we use nullification in different ways to suggest amino acids that may influence specificity in ligand binding cavities and protein-protein interfaces. Calibration of both nullification techniques is discussed in Supplemental Text S1.

2.4.1 Cavity fields. Amino acids that create electrostatic differences between two ligand binding cavities can cause different binding preferences. To identify amino acids like these, we begin with a *test* protein and a *reference* protein with different binding preferences and previously defined ligand binding cavities. First, we use ESBRI [78] to scan for intramolecular salt bridges. Second, we structurally align the test protein to the reference protein. Third, at an electrostatic threshold k selected by the user, we compute the cavity field of the reference structure, r, at k. Fourth, we systematically compute variants of the electrostatic field of t, where each variant exhibits a different nullified amino acid i. Once computed, the variant potential fields and the ligand binding cavity of the test structure are used to generate a variant cavity field t_i, for each nullified i, at isopotential k.

Each nullified cavity field t_i is then compared to the reference cavity field r by computing the volume of their CSG difference, $v(D_{t_i r})$. Based on this difference, we can propose several explanations for the impact of amino acid i on molecular recognition: If $v(D_{t_i r})$ is similar to $v(D_{tr})$, then nullifying i has little effect on the electrostatic differences between t and r, so we assume that i is not responsible for the differences in specificity between t and r. However, if $v(D_{t_i r})$ is significantly smaller than $v(D_{tr})$, then nullification of i reduces electrostatic differences between t and r. In this case, the original effect of i must have been to make the fields more different. Since electrostatic differences can be a sufficient reason for one protein to stabilize a binding partner that another cannot, we infer that i may be an electrostatic cause for different binding preferences.

Throughout this process we may observe that two amino acids i and j both appear, independently, to create significant differences between t and r. This observation, however, provides no information to compare their relative influence on specificity. Notably, even if $v(D_{t_i r}) - v(D_{tr})$ and $v(D_{t_j r}) - v(D_{tr})$ are different, they can also be differentially affected by other biophysical phenomena, and so we cannot infer that i affects specificity more or less than j does. We may also observe that $v(D_{t_i r})$ becomes greater than $v(D_{tr})$, indicating that nullification of i increases differences between t and r, suggesting that the effect of i is to make the electrostatic fields of t and r more similar. This observation may be true but it is insufficient to imply that i causes t and r to have similar binding preferences because other biophysical differences may prevent similar molecules from binding, despite the electrostatic similarities. Finally, if we observe that an amino acid i is part of an intramolecular salt bridge and that $v(D_{t_i r})$ is significantly smaller than $v(D_{tr})$, we infer that i is part of a salt bridge nearby the cavity and that mutating i would reduce cavity stability and alter electrostatic properties inside the cavity. By evaluating every amino acid in the manner above, VASP-E yields an electrostatic analysis relating each amino acid to its possible effect on binding.

Finally, we define a conservative prediction threshold for identifying amino acids that influence specificity. First, we compute $\mu = v(D_{tr})$, the volume of the difference between t and r without nullifications. μ represents the baseline electrostatic differences between the two cavity fields. Second, we find the amino acid i such that $\omega = v(D_{t_i r})$ is minimized for all i. ω represents the maximum degree to which the nullification of an amino acid can cause t and r to be similar. We define the prediction threshold $p = \mu - ((\mu - \omega)/2)$, which represents electrostatic differences reduced by one half of that achieved by ω. We predict that any amino acid j, where $v(D_{t_i r}) < p$, creates an electrostatic influence on specificity because nullifying it causes the cavity fields of t and r to become at least half as similar as is possible for any amino acid.

2.4.2 Interface fields. For protein-protein interfaces, we can perform a similar analysis to identify amino acids that affect electrostatic complementarity. Here, we begin with the structure of an input complex C and a user-selected threshold of electrostatic potential k. First, we use ESBRI [78] to scan for intramolecular salt bridges in each unit of the complex. We then generate interface fields at k and $-k$. Next, we create copies of the input complex, C_i, where one amino acid, i, is nullified. For each variant complex C_i, we generate interface fields at k and $-k$ as well.

Next, we compare the interface fields of each variant complex C_i to C, measuring the difference in electrostatic complementary $d(C_i, C)$ between the variant and input complexes. Using the value of the difference, we can draw several inferences about the nature

of electrostatic complementarity in the complex: If $d(C_i,C) \approx 0$ for some amino acid i, then nullifying i causes few differences in electrostatic complementarity at the protein-protein interface. We can thus infer that i has an insignificant electrostatic influence on affinity. If for some other i, $d(C_i,C)$ is significantly negative, then we infer that electrostatic complementarity is diminished in the variant complex relative to the input complex. i must therefore contribute to affinity when it is not nullified. Finally, if $d(C_i,C)$ is significantly positive, then we infer that electrostatic complementarity is enhanced by the nullification of i, implying that i is an electrostatic hindrance to binding affinity.

We may observe that nullification of two amino acids i and j independently result in significant changes to electrostatic complementarity. In such cases, the degree to which electrostatic complementarity is altered by i relative to j is insufficient to indicate their relative influence on affinity. We cannot draw this inference because other biophysical phenomena may unequally influence electrostatic complementarity for i and j, making their relative influence incomparable. We may also observe that an amino acid i is part of an intramolecular salt bridge and that nullification of i results in a significant change in electrostatic complementarity. In this case, we infer that i is part of a salt bridge nearby the interface and that mutating i would result in a destabilization of the protein at the interface and a reduction in binding affinity.

Finally, we define two conservative prediction thresholds to predict electrostatically influential amino acids in protein-protein interactions. Given a complex C to be evaluated at isopotential threshold k, we find the amino acid i such that $\omega = d(C,C_i)$ is minimized and the amino acid j such that $\Omega = d(C,C_j)$ is maximized. Amino acids i and j represent the amino acids that most increase and most decrease electrostatic complementarity in C upon nullification. We define the upper prediction threshold $P = \Omega/2$. If the nullification of an amino acid x increases electrostatic complementarity beyond P, then we predict that it reduces the electrostatic complementarity of the complex enough to reduce affinity. We also define a lower prediction threshold $p = \omega/2$. If the nullification of an amino acid y decreases electrostatic complementarity below p, we predict that this amino acid contributes to electrostatic complementarity of the complex enough to enhance affinity. In the case where amino acids at the interface do not act to increase electrostatic complementarity, we only apply the upper prediction threshold if fewer than 10% of the amino acids in the protein cause electrostatic complementarity to surpass P. We apply the same requirement to amino acids falling below p.

2.5 Clustering cavity fields based on volumetric similarity

Cavity fields based on a given family of proteins were clustered based on the Jaccard distance $J(x,y)$.

$$J(x,y) = 1 - \frac{v(x \cap y)}{v(x \cup y)}$$

where x and y are cavity fields, and $v(x \cap y)$ and $v(x \cup y)$ are the volume of the CSG intersection and CSG union of x and y, respectively. By normalizing the volume of the intersection by the volume of the union, the Jaccard distance permits cavity fields to be compared without bias relating to total volume. Cavity fields that have a low Jaccard distance have greater volumetric similarity than cavity fields with higher Jaccard distances. Using the neighbor program from Phylip [79], we summarized the pattern of volumetric similarities and variations between all pairs with UPGMA clustering (unweighted pair group method with arithmetic mean).

2.6 Clustering other measures of protein similarity

Members of a given family of proteins were also clustered based on amino acid sequence alignments and backbone structure alignments. ClustalW 2.0.7 was used to compute multiple sequence alignments. The resulting alignments were passed to the protpars tool from Phylip [79], to generate a maximum parsimony clustering of the protein sequences. Ska [80] was used to compute backbone structure alignments, which we used to generate a pairwise superposition of every structure onto a selected individual. The root mean squared distance (RMSD) between aligned alpha carbons was clustered via UPGMA, using the neighbor tool from Phylip [79]. Finally, Clustal Omega [81] was used to compute multiple sequence alignments and generate a neighbor joining tree.

2.7 Data set construction

Because VASP-E is designed to identify electrostatic influences on specificity, we validate it using families of proteins for which the mechanisms that achieve specificity are well understood and fundamentally electrostatic. The serine protease and cysteine protease superfamilies were selected for validating that VASP-E finds amino acids that influence protein-ligand binding specificity because many mutational studies confirm the role of specific residues in achieving specificity. The same studies permit the validation of VASP-E as a method for clustering proteins based on ligand binding preferences.

The protein data bank (PDB) [73] contains the structures of 681 serine proteases from the trypsin and chymotrypsin subfamilies, and 859 cysteine proteases from the cathepsin B, cathepsin L, and papain subfamilies. From each subfamily, we first removed mutants and functionally undocumented structures. Then we removed structures with greater than 90% sequence identity, creating a nonredundant subset of 12 serine proteases and 4 cysteine proteases. Filtering in this order maximized the number of diverse representative structures. Serine proteases averaged 51% sequence identity and cysteine proteases averaged 40% sequence identity.

We used ska [80] to structurally align the serine proteases to bovine chymotrypsin (pdb: 8gch) and the cysteine proteases to papaya papain (pdb: 1pad). Chymotrypsin and papain were selected because they are in complex with a peptide substrate. Using a method described earlier [37], substrate residues in the S1 subsite of the serine proteases and the S2 subsite of the cysteine proteases were used to generate a solid representation of the binding cavity in all structures. The binding cavity representation and the electrostatic field in each structure was then used to create cavity fields with the method in Section 2.2.

We demonstrate the comparison of interface fields on two protein complexes: barnase-barstar (pdb: 1brs) and rap1A-RAF (pdb: 1c1y). We selected these complexes because electrostatic potential is known to affect their binding preferences and because detailed experimental studies have established how binding preferences are affected by mutations on both sides of the interface. These studies create a well-defined gold standard for evaluating how accurately VASP-E can predict amino acids that alter binding preferences. The data set is summarized in Fig. 5.

2.8 Implementation details and performance

VASP-E was developed in ansi C/C++ using gcc (the Gnu Compiler Collection) version 4.4.7, on 64 bit linux-based computing platforms. Experimentation was performed on Corona,

Serine Protease superfamily:
Test family: Trypsins (EC 3.4.21.4)
2f91, 1fn8, 2eek, 1h4w, 1bzx, 1aq7, 1ane, 1aks, 1trn, 1a0j
Reference Structure: Chymotrypsins (EC 3.4.21.1)
8gch, 1eq9

Cysteine Protease Superfamily:
Test family: Cathepsin B (EC 3.4.22.1)
1csb, 1ito
Reference Structure: Cathepsin L (EC 3.4.22.5)
1icf
Reference Structure: Papain (EC 3.4.22.2)
1pad

Barnase/Barstar Complex: 1brs

Rap1a/raf Complex: 1c1y

Figure 5. PDB codes of structures used.
doi:10.1371/journal.pcbi.1003792.g005

a cluster at Lehigh University with 1056 Opteron cores (model 6128) running at 2.0 Ghz. Each compute node on corona had 16 cores with access to either 2 or 4 GB of random access memory (RAM) per core. VASP-E is a single-threaded process that runs on one core and approximately 1 GB of random access memory. All experimentation was conducted at .5 Å resolution, which permitted accurate results and practical runtimes.

Visualization for some figures was performed with SURFview, a tool written using the OpenGL library and running on Intel Core i7 and Nvidia Geforce GTX 660 chipsets, in Microsoft Windows 7. Trees representing clusterings were visualized using Newick Utilities [82].

The performance of VASP-E depends on the volume and resolution of the molecular surfaces or electrostatic isopotentials analyzed. On our dataset, generating solid isopotentials for entire proteins required approximately 9.5 seconds on average, to process an average of 1,337,083 lattice cubes. Comparing cavity fields required 1.06 seconds on average, to process an average of 41,984 cubes via CSG, while interface fields from two complexes required 23.4 seconds on average, to process an average of 729,321 cubes.

The website http://www.cse.lehigh.edu/~chen/software.htm hosts the software and primary data associated with this paper for public download.

Results

3.9 Serine proteases

Serine proteases exhibit affinity for amino acids at specificity subsites called S4, S3, …, S1, S1′, …, S3′, S4′ [83]. Each subsite recognizes substrate residues P4, P3, …, P1, P1′, …, P3′, P4′, enabling the protease to selectively cleave the peptide bond between P1 and P1′. Trypsins are digestive serine proteases that narrowly prefer positively charged amino acids [84] at *P*1. Their selectivity is assisted by the strongly negative electrostatic character of S1. In contrast, chymotrypsins hydrolyze peptide bonds following large hydrophobic amino acids [85] and exhibit considerably less electrostatic potential at their S1 subsite.

Using VASP-E, we identified amino acids that create electrostatic differences between trypsins and chymotrypsins at S1. Fig. 6 reports the average volumetric difference between cavity fields from all trypsins in our dataset and the cavity field of bovine chymotrypsin (pdb: 8gch), where each trypsin residue has been nullified individually. Volume differences were computed for cavity fields generated at −2.5, −5.0, −7.5, and −10.0 kT/e. Volumetric differences between nullified trypsin and chymotrypsin cavity fields varied most at −10 kT/e, so a prediction threshold was computed for differences at this level. The average volumetric difference between trypsin and chymotrypsin cavity fields remained nearly constant for almost all residue nullifications and all four thresholds. Nullifying almost all trypsin residues does not make the very different electrostatic environments of the trypsin and chymotrypsin S1 pockets more similar.

One notable exception stands out. Nullifying aspartate 189 in all trypsins results in a large reduction in the average electrostatic difference with chymotrypsin at all potential thresholds, suggesting that the presence of aspartate 189 makes their S1 pockets electrostatically different. Fig. 7 illustrates the effect that nullifying aspartate 189 has on the electrostatic difference between chymotrypsin and trypsin, using bovine chymotrypsin and atlantic salmon trypsin as examples. VASP-E examines only the volumetric intersection of their S1 cavities, where the binding cavities have no steric differences (Fig. 7b). In unmodified trypsin (Fig. 7d), the intersection region exhibits a 152 $Å^3$ region with electrostatic potential less than or equal to −10 kT/e. Once D189 is nullified, the region with potential less than or equal to −10 kT/e drops to 32 $Å^3$ (Fig. 7e). In comparison, regions of negative electrostatic potential in chymotrypsin, where the S1 cavity overlaps with that of trypsin, is small and remains small when S189 is nullified (Fig. 7f,g). Similar effects were observed with other trypsins. These indications predict experimentally established observations that the negatively charged aspartate 189, at the bottom of the S1 pocket, creates the specificity of trypsin for positively charged amino acids [48,86].

Fig. 8 illustrates a UPGMA clustering of cavity fields from trypsin and chymotrypsin S1 cavities, generated at −10 kT/e. The topology of the tree, which reflects electrostatic similarities and differences measured with the Jaccard distance, correctly separated the chymotrypsins as outliers from the trypsins. This result indicates that the electrostatic characteristics measured by VASP-E correlate with similarities and differences in serine protease binding preferences.

Clusterings based on cavity fields generated at −2.5, −5.0, −7.5, or −10.0 kT/e (Fig. S1) illustrate that the classification is correct for a range of isopotential thresholds, though they becomes less accurate as thresholds approach neutral charges. Also, relative to comparisons of protein sequences and structures, separated trypsins and chymotrypsins less well (Fig. S2).

3.10 Cysteine proteases

Cathepsin B is involved in the onset of pancreatitis [87] and the malignant progress of tumors [88]. Following the same subsite/substrate numbering scheme as serine proteases, Cathepsin B cleaves a peptide bond following two positively charged amino acids that bind in its S1 and S2 subsites [89]. The S2 subsite exhibits a strong negative potential that enables the recognition of positively charged side chains. In contrast, cathepsin L and papain prefer bulky hydrophobic amino acids at *P*2 [90,91], and both exhibit an uncharged S2 subsite [91].

We used VASP-E to identify amino acids that create electrostatic differences between cathepsin B and cathepsin L. Fig. 9 illustrates the average volumetric difference between cavity fields representing S2 in cathepsin B and human cathepsin L (pdb: 1icf) generated at −2.5, −5.0, −7.5, and −10.0 kT/e. Volumetric differences between cavity fields with different nullified amino acids were greatest at −2.5 kT/e, so a prediction threshold was computed for differences at this level. The average volumetric difference was nearly constant for almost all residue nullifications. Insignificant fluctuations in the volumetric difference were approximately the same magnitude as in serine proteases.

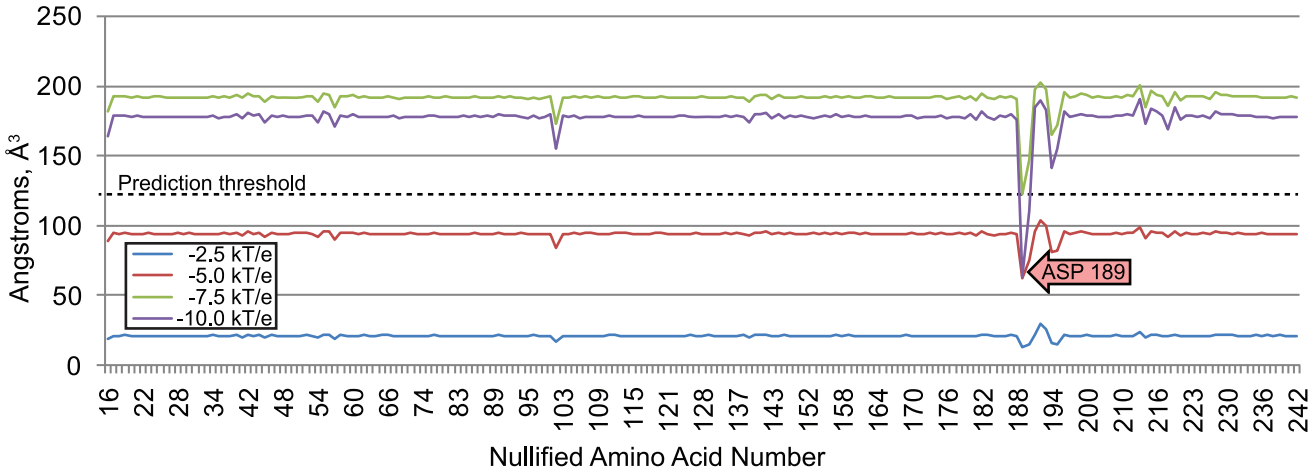

Figure 6. Average volume differences between chymotrypsin and trypsin cavity fields with nullified amino acids. The red arrow indicates a trypsin residue associated with increased electrostatic similarity (downward spikes) when it is nullified. The dashed line represents the average prediction threshold between chymotrypsin and trypsin cavity fields.
doi:10.1371/journal.pcbi.1003792.g006

The nullification of two amino acids, glutamic acid 171 or glutamic acid 245, reduced electrostatic differences between cathepsin B and cathepsin L beyond the prediction threshold. This observation suggests that both amino acids create electrostatic differences between the S2 subsites of cathepsin B and L. Indeed, glutamic acid 245 has been shown to cause Cathepsin B to bind arginine residues at the S2 cavity [90], while cathepsin L prefers phenylalanines. In glutamic acid 171, one of the carboxylate oxygens is involved in a hydrogen bond and the other is free to form other interactions in the S2 pocket. Such interactions have been observed with positively charged inhibitors [92,93], again in contrast with cathepsin L.

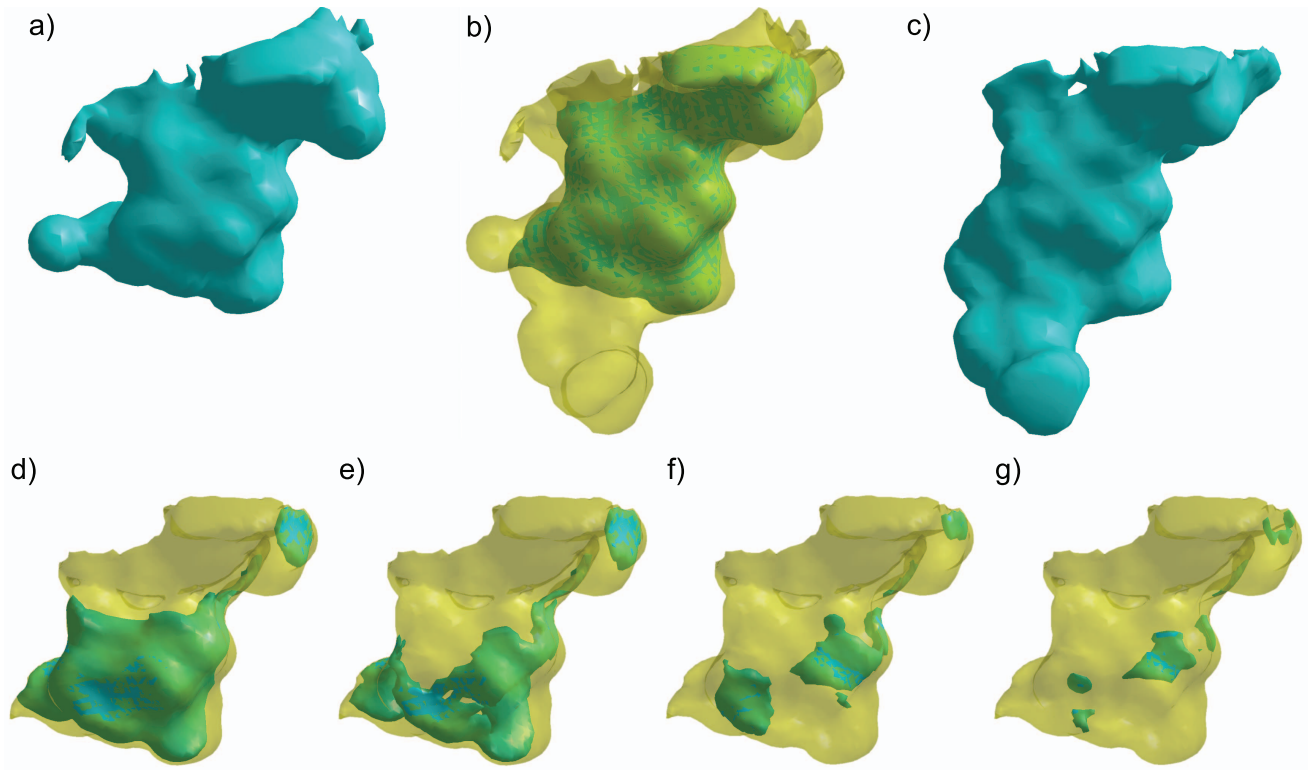

Figure 7. A visual examination of the nullification of aspartate 189 of trypsin. a) S1 cavity of atlantic salmon trypsin (pdb: 1a0j) shown in teal. b) Intersection region (teal) of S1 cavities from trypsin and chymotrypsin (transparent yellow). c) S1 cavity of bovine chymotrypsin (pdb: 8gch) shown in teal. Inset figs. d-g illustrate cavity fields, all with potential less than −10 kT/e (teal), inside the intersection region (transparent yellow). d) The wild type trypsin cavity field occupies 152 Å^3. e) The trypsin cavity field with D189 nullified (32 Å^3). f) The wild type chymotrypsin cavity field (9 Å^3). g) The chymotrypsin cavity field with D189 nullified (2 Å^3).
doi:10.1371/journal.pcbi.1003792.g007

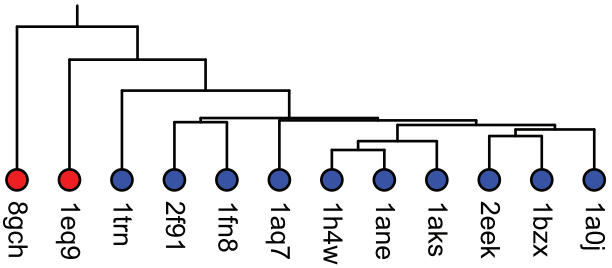

Figure 8. Patterns of electrostatic similarity in the S1 specificity pockets of trypsins and chymotrypsins, relative to P1 binding preferences. The color coding, which is independent of tree topology, indicates the types of P1 residue preferred by each protein. Trypsins (blue) prefer basic amino acids and chymotrypsins (red) prefer large hydrophobic amino acids. The topology of the tree reflects patterns of similarity measured with the Jaccard distance. Proteins on adjacent branches have greater similarity than proteins on different subtrees. The topological separation of the chymotrypsins from the trypsins indicates that similarities and differences in the electrostatic character of S1 subsites, which create the differences in their binding preferences, were detected and correctly classified by VASP-E, using the Jaccard distance.
doi:10.1371/journal.pcbi.1003792.g008

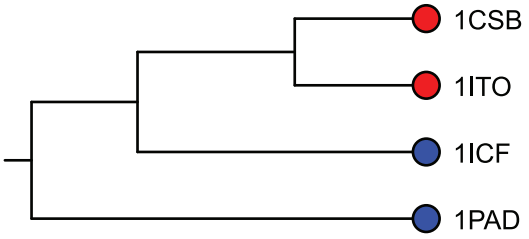

Figure 10. Patterns of electrostatic similarity in the S2 specificity pockets of cathepsin B, cathepsin L, and papain. The color coding, which is independent of tree topology, indicates the types of P2 residue preferred by each protein. Cathepsin B's (red) prefer basic amino acids and cathepsin L and papain (blue) prefer large hydrophobic amino acids. The topology of the tree reflects patterns of similarity measured with different comparison algorithms. Proteins on adjacent branches have greater similarity than proteins on different subtrees. The topological separation of the cathepsin B's from cathepsin L and papain indicates that similarities and differences in the electrostatic character of S2 subsites, which create the differences in their binding preferences, were detected and correctly classified by VASP-E, using the Jaccard distance.
doi:10.1371/journal.pcbi.1003792.g010

Fig. 10 plots a UPGMA clustering of cavity fields based on S2 subsites from cysteine proteases in our data set. Cavity fields were generated at −2.5 kT/e. The topology of the tree describes electrostatic similarities and differences measured with the Jaccard distance. It is apparent that the tree structure clusters cathepsin B cavities, setting them apart from those of cathepsin L and papain, which have different binding preferences. Cavity fields produced at −2.5, −5.0, −7.5, and −10.0 kT/e, shown in Fig. S3, cluster in a similar manner. This pattern of separations demonstrates that VASP-E is correctly identifying electrostatic conservations and variations that correlate to binding preferences in the S2 subsite. Global sequence and structure alignments separated the cysteine proteases as well as the Jaccard distance clustering (Fig. S4).

3.11 Barnase-barstar

Barnase is an guanine-preferring endo-ribonuclease expressed by Bacillus amyloliquefaciens [94] whose activity, without inhibition by barstar, can be lethal to the cell [95]. Barstar inhibits barnase by forming an extremely tight complex with close steric and electrostatic complementarity at many amino acids across the binding site [96]. We used VASP-E to identify mutations that enhance or diminish electrostatic complementarity.

3.11.1 Nullifications of barnase amino acids. Fig. 11 illustrates comparisons of wildtype and modified barnase-barstar interface fields, where nullifications were performed on Barnase. The magnitude of volume changes observed were much larger than in the cavity fields examined earlier because the volume of the interface is much larger than the cavities. We evaluated the impact of nullifying individual amino acids with values of k equal to 1.0, 3.0, 5.0, 7.0, and 9.0 kT/e. Differences in electrostatic complementarity caused by nullifying some amino acids were greatest at $k = 1.0$ kT/e, so this value of k was used to set upper and lower prediction thresholds. Nullification of most barnase amino acids resulted in small changes in electrostatic complemen-

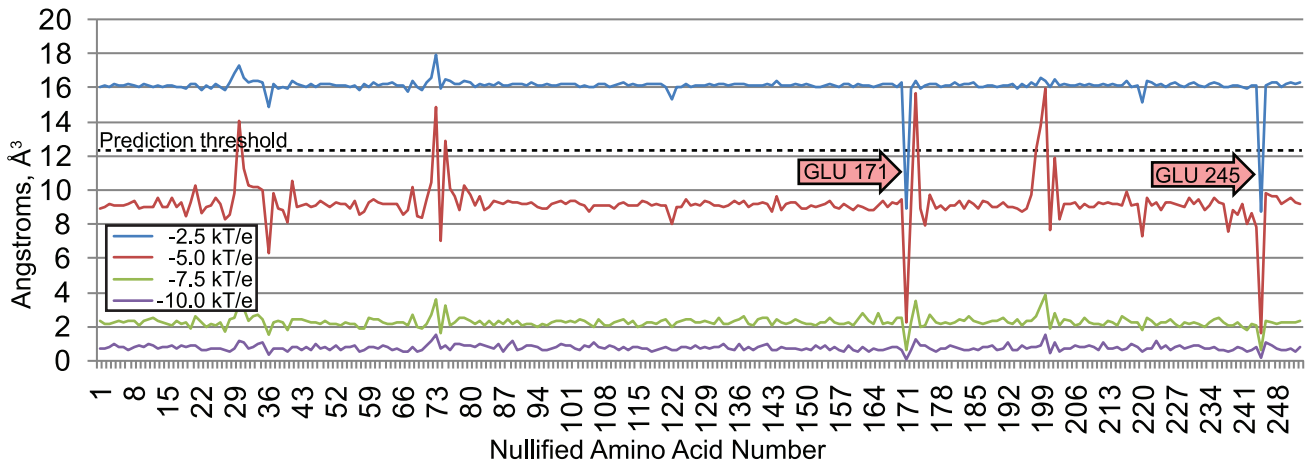

Figure 9. Average volume differences between cathepsin L and cathepsin B cavity fields with nullified amino acids. The red arrows indicate amino acids in cathepsin B associated with increased electrostatic similarity (downward spikes) to cathepsin L, when they are nullified.
doi:10.1371/journal.pcbi.1003792.g009

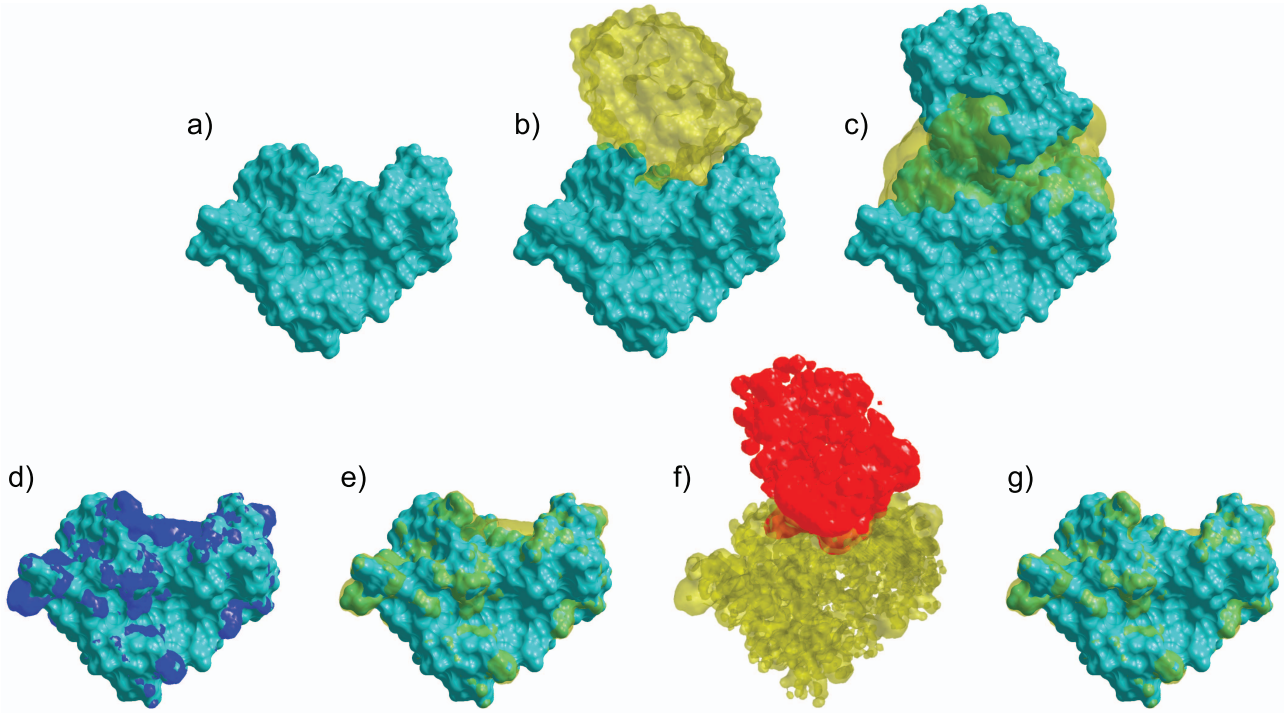

Figure 11. Volume differences between interface fields of wildtype barnase/barstar and a barnase/barstar complex with electrostatic nullifications in the barnase residues. The red arrows indicate amino acids in barnase that are associated with decreased electrostatic complementarity with barstar, when they are nullified. Blue arrows indicate amino acids associated with increased electrostatic complementarity, when they are nullified. Green arrows indicate amino acids below the prediction threshold that are known to influence specificity.
doi:10.1371/journal.pcbi.1003792.g011

tarity below both prediction thresholds. However, nullification of a few amino acids created very large increases and decreases in complementarity between wildtype and modified interface fields.

For four barnase residues, K27, R59, R83, and R87, nullification significantly reduced electrostatic complementarity, predicting correctly that mutations abrogating net charge at these positions could reduce affinity. These predictions are consistent with experimental observations established earlier: K27A decreases association rates by a factor of 7 to 10 times [97–99]. R59A reduces association rates by a factor of 7 to 10 times [98,99]. R83A decreases association rates by 4 to 6 fold [95,97,99]. R87A decreases association rates by 2 to 3 fold [97–99].

Nullification of barnase residues 54 and 73 significantly increased electrostatic complementarity, correctly predicting that substituting these amino acids with alanine should increase affinity. Predictions for D54 and E73 reproduced established observations: Substituted individually, D54A and E73A increase association rates by 2 to 4 fold [98,99]. Also, D75 is involved in an intramolecular salt bridge, and is thus predicted to reduce the stability of barnase and it's ability to form a complex with barstar. This prediction is correct; the substitution of D75 with asparagine, a nearly isosteric but uncharged analogue of aspartate, is known to diminish complex stability by 4.80 kcal/mol [100].

Three known influences on affinity fell below our prediction threshold. Nullifying residues 39 and 102 reduced electrostatic complementarity, but not significantly enough to achieve our prediction threshold. The mutation K39A is known to reduce affinity [95], and the mutation H102A reduces association rates less than 2 fold [97–99]. Also, it is known that replacing glutamic acid 60 with alanine is known to increase association rates by 2 to 4 fold [98,99], but nullifying glutamic acid 60 did not generate a large increase in electrostatic complementarity. While VASP-E made no incorrect predictions, the application of a conservative prediction threshold caused some influential amino acids to be missed.

Nullifications of influential amino acids identified by VASP-E create changes in electrostatic complementarity that can be localized to specific regions. For example, Fig. 12 illustrates the effect of nullifying lysine 27 in barnase. In the interface region, lysine 27 is responsible for a large positively charged region of electrostatic potential that extends outwards towards barstar (Fig. 12e). This region overlaps considerably with a negative isopotential from barstar (Fig. 12f). When lysine 27 is nullified, the positively charged isopotential on barnase collapses (Fig. 12g), and electrostatic complementarity is substantially reduced. This ability to identify spatial regions of electrostatic complementarity, and thus provide insights into the mechanisms that control specificity, is unique to VASP-E.

3.11.2 Nullifications of barstar amino acids. Fig. 13 plots comparisons of wildtype and modified Barnase-Barstar interface fields, where nullifications were performed on Barstar. Because Barstar is interacting with Barnase, the same calibration threshold, $k = 1.0\ kT/e$, was used. Also, because more than 10% of amino acids in barstar are above the threshold, an upper prediction threshold was not used, suggesting that there are no outliers on the positive end. Nullifications of several amino acids created distinctive differences between wildtype and modified interface fields.

Nullifying three barstar residues, 35, 39 and 80 reduced electrostatic complementarity. These observations correctly predict experimental observations that these amino acids are crucial for affinity between barnase and barstar, and that diminishing their electrostatic contribution interferes with binding: Charge reversal mutations individually converting aspartate 35 and 39 to

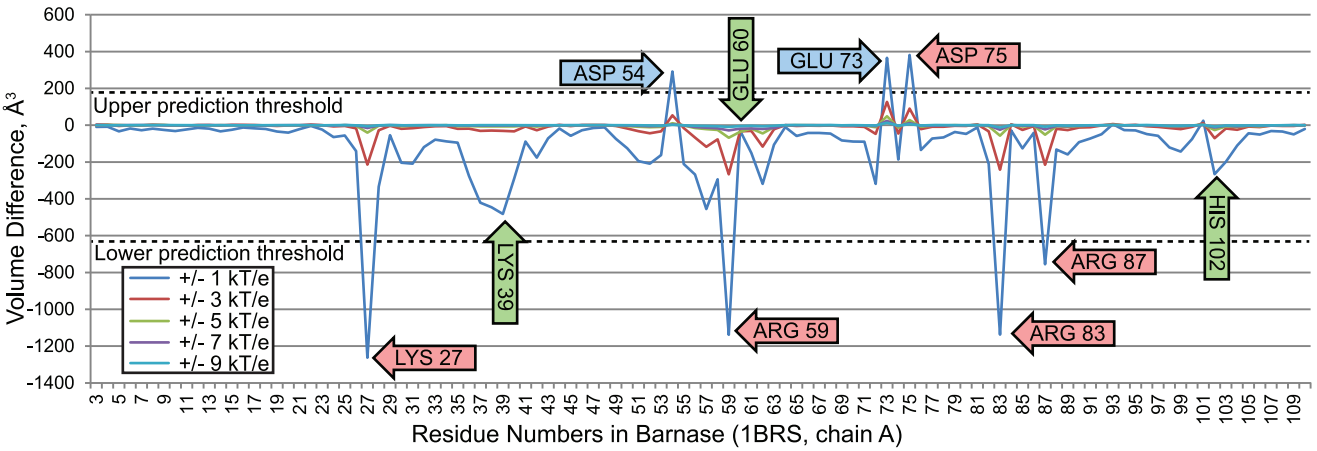

Figure 12. A visual examination of the nullification of lysine 27 in barnase. a) The molecular surface of Bacillus amyloliquefaciens barnase (teal). b) The molecular surface of Bacillus amyloliquefaciens barstar (transparent yellow) and barnase (teal). c) The interface region (transparent yellow) between barnase and barstar (teal). d) Electrostatic isopotential at +3 kT/e (blue) near barnase (teal). e) The same isopotential shown in transparent yellow. f) The electrostatic isopotential at −3 kT/e near barstar (red) and it's overlap with the electrostatic isopotential at +3 kT/e near barnase (transparent yellow). g) Electrostatic isopotential at +3 kT/e (blue) near barnase (teal), where Lysine 27 is nullified.
doi:10.1371/journal.pcbi.1003792.g012

lysine were shown to halt the inhibition of barnase by barstar [99,101]. Mutation of glutamic acid 80 to alanine reduces the binding energy by .5 kcal/mol and increases the dissociation constant by 2.5 fold relative to the wildtype complex.

The nullification of glutamic acid 76 insufficiently reduced electrostatic complementarity to be associated with a prediction. Nonetheless, the mutation of E76 to alanine was shown to reduce the binding energy by 1.6 kcal/mol and increases the dissociation constant by 10 fold relative to the wildtype complex [99]. As was the case with barnase, VASP-E made no incorrect predictions but the application of a conservative prediction threshold caused some influential amino acids to be missed.

Nullifying the uncharged interfacial amino acids 29–31, 36–38, and 40–46 generated increases in electrostatic complementarity via electrostatic focusing. This enhancement creates isopotentials with larger volume, especially when the isopotentials are generated at low absolute thresholds (e.g.+/− 1 kT/e). Since these amino acids are uncharged, their nullification enlarges the isopotentials of nearby charged amino acids D35 and D39.

3.12 Rap1a-Raf

Ras is a master regulator that transmits a wide range of signals via protein-protein interactions. Downstream, its effectors are involved in many crucial systems, including cell cycle progression, cell division, apoptosis, lipid metabolism, DNA synthesis, and cytoskeletal organization [102–104]. While the structure of ras in complex with these effectors is unknown, rap1a, a homolog of ras (>50% sequence identity), can serve as a substitute. Rap1a has an essentially identical binding interface and binds competitively with the same downstream effectors [105], such as raf, an oncogene involved in ERK 1/2 signaling [106]. Here, we use VASP-E to

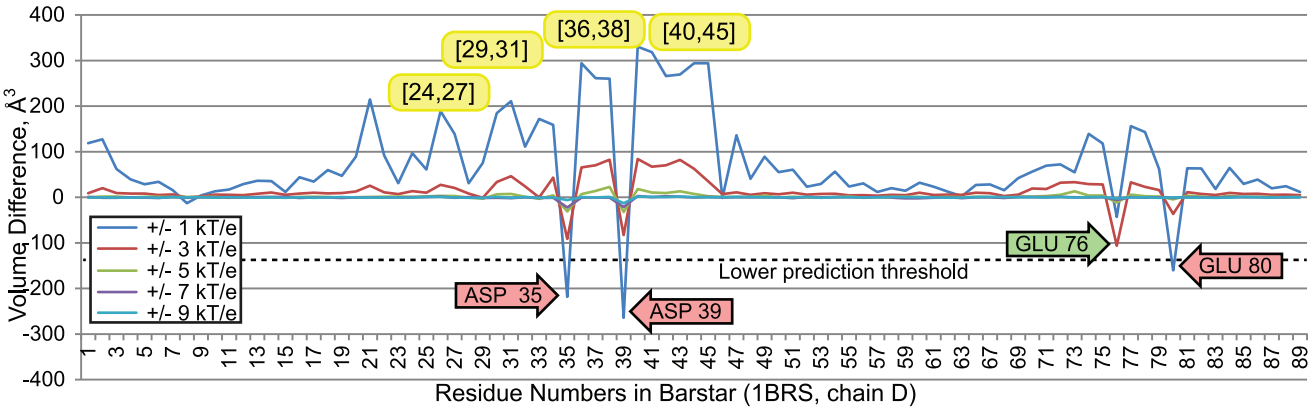

Figure 13. Volume differences between interface fields of wildtype barnase/barstar and a barnase/barstar complex with electrostatic nullifications in the barstar residues. The red arrows indicate amino acids in barstar that are associated with decreased electrostatic complementarity with barstar, when they are nullified. Numbers in yellow ovals indicate inclusive intervals of amino acids where electrostatic focusing enhances the volume of the electrostatic potential inside the barnase/barstar interface. The green arrow indicates an amino acid below the prediction threshold that is known to influence specificity.
doi:10.1371/journal.pcbi.1003792.g013

examine the effect of charge nullification on the rap1a-raf interface to make predictions on the effect of mutation on ras-raf binding.

3.12.1 Nullifications of Rap1A amino acids. Fig. 14 plots comparisons of wildtype and modified Rap1A-Raf interface fields, where nullifications were performed on Rap1a. We evaluated the impact of nullifying individual amino acids with values of k equal to 1.0, 3.0, 5.0, 7.0, and 9.0 kT/e. Differences in electrostatic complementarity caused by nullification were greatest at $k=1.0$ kT/e, so this value of k was used to set upper and lower prediction thresholds. Nullification of most barnase amino acids resulted in small changes in electrostatic complementarity below both prediction thresholds. However, nullification of several amino acids created very large increases and decreases in complementarity between wildtype and modified interface fields.

Nullification of six rap1a residues, 33, 37, 38, 54, 57 and 62 reduced electrostatic complementarity beyond the lower prediction threshold, suggesting that loss of charge mutations would reduce complex affinity. These predictions were consistent with established experimental observations: Substituting aspartate 33 for alanine in rap1a results in a binding energy reduction of 1.2 kcal/mol [107]. Glutamic acid 37, in both rap1a and ras, forms hydrogen bonds with with R59 and R67 in raf [105]. Substituting E37 with glycine would break these bonds, and in ras-raf, E37G inhibits the formation of the complex [108]. Substituting aspartate 38 for alanine in ras, eliminating its contribution to electrostatic complementarity and removing a hydrogen bond, reduces its rate of association with raf by 72 fold [107,109]. Glutamic acid 54 forms a hydrogen bond with arginine 67 of raf. Mutations of R67 that break this bond reduce the rate of association by 12 fold [108,110]. Substituting E54 with alanine would break the same bond and likely achieve a similar effect. Substituting aspartate 57 for alanine in Ha-Ras causes a total loss of affinity to raf [111]. Finally, glutamic acid 62 is a conserved amino acid that radically affects binding in a range of RAS homologs when mutated [112].

Nullification of two rap1a residues, 31 and 41, increased electrostatic complementarity beyond the upper prediction threshold, suggesting that mutations removing their net charge should also increase affinity. Established results confirm these observations: Charge reversal of lysine 31 to glutamic acid is known to create a 30 fold increase in affinity [110]. In Ha-ras, a substitution of arginine 41 to alanine is known to increase affinity by 11.8 fold [111,113].

Finally, VASP-E predicted that the nullification of lysine 5 could result in an increase in binding affinity. This observation suggests that lysine 5 may normally reduce affinity. However, to our knowledge, no current experimental results that establish this claim, and hence we leave it as an open prediction.

3.12.2 Nullifications of Raf amino acids. Fig. 15 plots comparisons of wildtype and modified Rap1A-Raf interface fields, where nullifications were performed on residues in raf. Because Raf is interacting with Rap1a, the same calibration threshold, $k=1.0$kT/e, was used. Also, because more than 10% of amino acids in Raf are above the threshold, an upper prediction threshold was not used, suggesting that there are no outliers on the positive end. Nullifications of several amino acids created distinctive differences between wildtype and modified interface fields.

Nullification of four residues in raf, 59, 67, 84, and 89 reduced electrostatic complementarity below the lower prediction threshold and correctly predicted experimentally established substitutions that correspond to reductions in affinity. Substituting arginine 59 with alanine is known to reduce the rate of association by 25 fold [108,110]. Substituting arginine 67 with alanine is known to reduce the rate of association by 12 fold [108,110]. Also, both arginine 59 and 67 form hydrogen bonds with glutamic acid 37 in rap1a. Loss of these hydrogen bonds inhibits complex formation [108]. Substitution of lysine 84 with alanine produces a 9.4 fold reduction in the association rate [110], and substitution of arginine 89 with leucine inhibits complex formation [105].

Nullification of lysine 65 did not reduce electrostatic complementarity below the lower threshold. While lysine 65 was therefore not predicted to have a significant electrostatic influence on specificity, the mutation K65A is known to reduce the rate of association by 4.5 fold [110]. While VASP-E made no incorrect predictions, the conservative prediction threshold caused K65 to be overlooked.

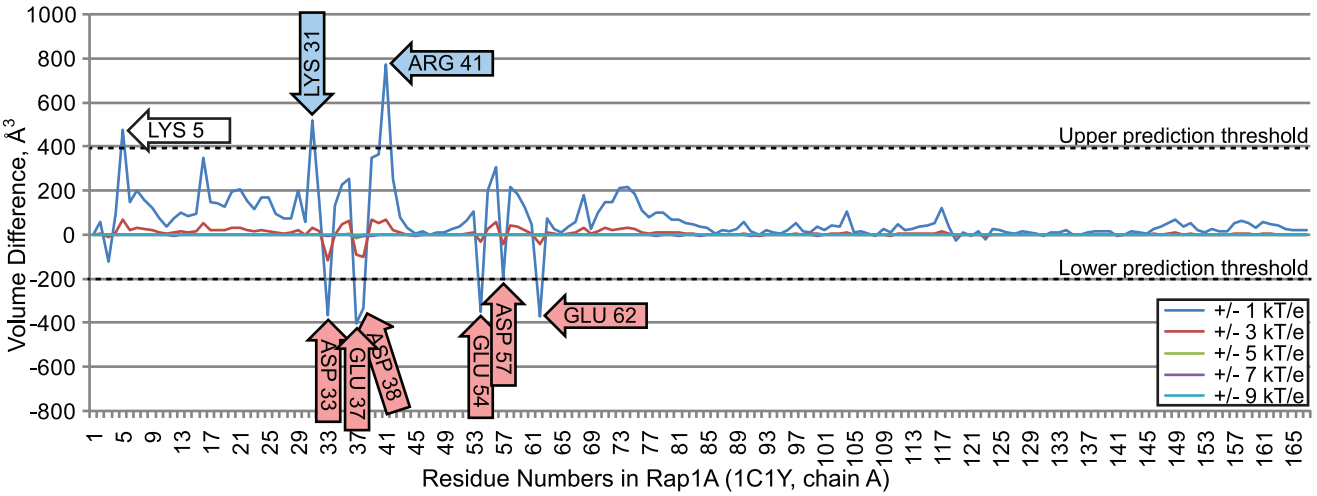

Figure 14. Volume differences between the interface fields of a wildtype rap1a/raf complex and a rap1a/raf complex with electrostatic nullifications in rap1a residues. The red arrows indicate amino acids in rap1a that are associated with decreased electrostatic complementarity with barstar when they are nullified. Blue arrows indicate amino acids associated with increased electrostatic complementarity, when they are nullified. The white arrow indicates an open prediction.
doi:10.1371/journal.pcbi.1003792.g014

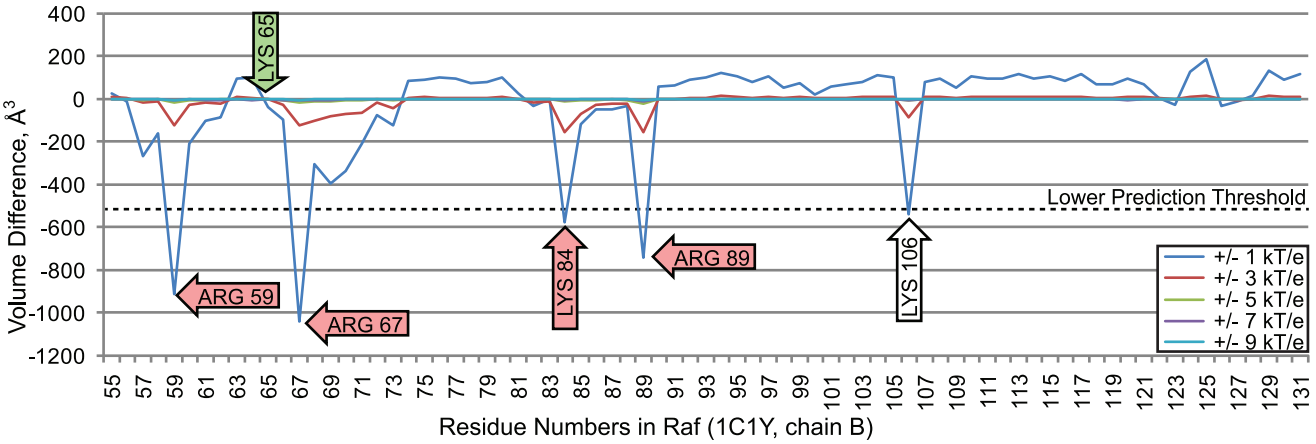

Figure 15. Volume differences between the interface fields of a wildtype rap1a/raf complex and a rap1a/raf complex with electrostatic nullifications in raf residues. The red arrows indicate amino acids in rap1a that are associated with decreased electrostatic complementarity with barstar when they are nullified. The green arrow indicates an amino acid below the prediction threshold that is known to influence specificity. The white arrow indicates an open prediction.
doi:10.1371/journal.pcbi.1003792.g015

3.13 Analysis of prediction performance on individual amino acids

By collecting the predictions made on our dataset, we can measure the prediction performance of VASP-E. We begin by counting true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs). TPs are defined as amino acids that are both predicted by VASP-E to have an influence on specificity and also published in experimental findings to have such an influence. The predictions detailed earlier in this section cite these findings as specific validation for the predictions made with VASP-E. FPs are amino acids that are both predicted by VASP-E to have an influence on specificity and are documented in the literature to not have an effect on specificity. TNs are amino acids that are predicted to not have an influence on specificity that are also documented in the literature to not have an effect on specificity. FNs are amino acids predicted to not have an influence on specificity but are established in the literature as having a role in specificity. Finally, we VASP-E made two predictions that were neither confirmed nor denied in the literature. We leave these two observations as open predictions and do not include them in our evaluation of prediction performance.

Of these statistics, TNs cannot be fully counted because no studies categorically classify the role of every amino acid in specificity, including those that are distant from the binding site. For this reason, we describe the number of true negatives as unknown. Nonetheless, we do not require TNs in order to compute *precision* and *recall*, two fundamental statistics used to evaluate the accuracy of a predictor. Precision is the fraction of predictions that are verified in experimental studies and recall is the fraction of verified experimental results that are correctly predicted. Using our conservatively defined prediction thresholds, every prediction made with VASP-E was verified, giving perfect precision, and most verified results were correctly predicted, giving strong recall. Precision and recall are reported together in Fig. 16.

Discussion

We have presented VASP-E, a new program for the comparison of electrostatic isopotentials. To our knowledge, VASP-E is the first program capable of comparing isopotentials using CSG, enabling a new unified approach to the characterization of protein-ligand and protein-protein binding specificity. In an application to the serine and cysteine proteases, we demonstrate that VASP-E is capable of reproducing known ligand binding preferences and of detecting differences in electrostatic potential among proteins that, based on global sequence and structure similarity, might have been expected to be similar. Subtle differences like these, which can arise from variations in single amino acids, can still be detected by VASP-E because they are reflected in differently shaped isopotentials.

Central to our approach is a novel solid representation of electrostatic isopotentials that can also represent regions within molecular surfaces. This seamless integration of two nearly orthogonal aspects of protein structure enables analytical capabilities that were not possible before. One capability is the identification of amino acids that create differences in electrostatic isopotentials at binding cavities. Using the molecular surface to exclude electrostatic variations outside the binding cavity, we identified three amino acids in trypsin and cathepsin B that create electrostatic differences in binding specificity. These predictions correctly reflected experimentally established observations regarding their electrostatic influence. VASP-E also finds amino acids that change electrostatic complementarity in protein-protein interfaces. In an analysis of the barnase-barstar and rap1a/raf

	True	False
Positive	25	0
Negative	Unknown	5

Open predictions: 2

$$\text{Precision: } \frac{TP}{TP+FP} = 100\%$$

$$\text{Recall: } \frac{TP}{TP+FP} = 83.3\%$$

Figure 16. Precision and recall performance of VASP-E.
doi:10.1371/journal.pcbi.1003792.g016

complexes, VASP-E predicted 22 amino acids that either increase or decrease affinity upon mutation, all in agreement with established experimental results. Solid representations enable a deconstructive analysis of electrostatic fields that permits the discovery of individual residues that influence binding preferences in protein-ligand and protein-protein binding sites.

As the first approach to the comparison of electrostatic isopotentials with CSG, VASP-E exhibits novel potential for useful experimental applications. In experimental settings, identifying mutants that may alter binding specificity can be a time consuming and expensive effort with many possible mutants to consider. VASP-E identifies amino acids that might play a role in specificity, and, in addition, it suggests a biophysical mechanism for that amino acid: It may increase or decrease electrostatic complementarity. This additional information, beyond simply identifying an important amino acid, provides utility beyond the identification of important amino acids because it suggests how that amino acid might be tested, such as by mutation to an uncharged or oppositely charged residue. When comparing protein-ligand binding cavities, pointing out amino acids that create electrostatic differences can inform experimental design.

VASP-E has the potential to serve broad applications. For example, identifying groups of amino acids that work together to achieve specificity can be an especially difficult problem, because of the combinatorial space of variants that must be considered. Nullification, as applied to individual amino acids in this paper, could be exhaustively applied to many combinations of residues to assist in experimental design. Given the rapid performance of VASP-E and the expanding availability of parallel computing, examining combinations of influential amino acids would also be very practical. Furthermore, the analysis of influential amino acids at protein-protein interfaces is not limited to dimers; the approach described here could be logically extended to higher order interactions. For such applications, interfaces between specific chains could be considered individually or in groups, to reflect the order in which the complex associates. Finally, while VASP-E is designed to identify subtle variations among highly similar proteins, VASP-E could in principle be used to analyze electrostatic similarities and differences among binding sites from very different proteins, as long as structural alignments could be correctly generated and binding cavities can be properly defined. These diverse applications suggest that the integrated representation and comparison of structure and electrostatics may offer an important new tool in the study of drug resistance and algorithms for specificity annotation.

Supporting Information

Figure S1 Patterns of electrostatic similarity in the S1 specificity pockets of trypsins and chymotrypsins, relative to P1 binding preference. The color coding in all trees, which is independent of tree topology, indicates the types of P1 residue preferred by each protein. Trypsins (blue) prefer basic amino acids and chymotrypsins prefer large hydrophobic amino acids (red). The topology of each tree reflects patterns of similarity measured with the Jaccard distance on cavity fields generated at different isopotential thresholds. In each tree, proteins on adjacent branches have greater similarity than proteins on different subtrees. The topology of the trees reflect UPGMA clustering of serine protease cavity fields generated at (a) 2.5 kT/e, (b) 5.0 kT/e, (c) 7.5 kT/e, (d) and 10.0 kT/e.
(EPS)

Figure S2 Patterns of similarity and variation in the sequence, backbone structure, and cavity fields of trypsins and chymotrypsin, relative to P1 binding preference. The color coding in all trees, which is independent of tree topology, indicates the types of P1 residue preferred by each protein. Trypsins (blue) prefer basic amino acids and chymotrypsins prefer large hydrophobic amino acids (red). The topology of each tree reflects patterns of similarity measured with different comparison algorithms. In each tree, proteins on adjacent branches have greater similarity than proteins on different subtrees. The topology of tree (a) reflects sequence similarity measured with Clustalw 2.0.7, the topology of (b) reflects backbone structure similarity measured with ska, the topology of (c) reflects cavity field similarity measured with the Jaccard distance, and the topology of (d) reflects sequence similarity as measured with clustal omega. Jaccard similarity positions serine proteases with similar P1 binding preferences more closely than the other similarity measures do.
(EPS)

Figure S3 Patterns of electrostatic similarity in the S2 specificity pockets of cathepsin B, cathepsin L, and papain, relative to P2 binding preference. The color coding in all trees, which is independent of tree topology, indicates the types of P2 residue preferred by each protein. Cathepsin B's (red) prefer basic amino acids and cathepsin L and papain (blue) prefer large hydrophobic amino acids. The topology of each tree reflects patterns of similarity measured with the Jaccard distance on cavity fields generated at different isopotential thresholds. In each tree, proteins on adjacent branches have greater similarity than proteins on different subtrees. The topology of the trees reflect UPGMA clustering of cysteine protease cavity fields generated at (a) 2.5 kT/e, (b) 5.0 kT/e, (c) 7.5 kT/e, (d) and 10.0 kT/e.
(EPS)

Figure S4 Patterns of similarity and variation in the sequence, backbone structure, and cavity fields of cysteine proteases, relative to P2 binding preference. The color coding in all trees, which is independent of tree topology, indicates the types of P2 residue preferred by each protein. Cathepsin B's (red) prefer basic amino acids and cathepsin L and papain (blue) prefer large hydrophobic amino acids. The topology of each tree reflects patterns of similarity measured with different comparison algorithms. In each tree, proteins on adjacent branches have greater similarity than proteins on different subtrees. The topology of tree (a) reflects sequence similarity measured with Clustalw 2.0.7, the topology of (b) reflects backbone structure similarity measured with ska, the topology of (c) reflects cavity field similarity measured with the Jaccard distance, and the topology of (d) reflects sequence similarity as measured with clustal omega. Jaccard similarity positions cysteine proteases with similar P2 binding preferences in a manner similar to the other similarity measures.
(EPS)

Text S1 Text S1 includes three supplemental notes describing the calibration of the method.
(PDF)

Acknowledgments

The author sincerely thanks Barry Honig for his thoughtful advice on this study and Remo Rohs for insightful discussions.

Author Contributions

Conceived and designed the experiments: BYC. Performed the experiments: BYC. Analyzed the data: BYC. Contributed reagents/materials/analysis tools: BYC. Wrote the paper: BYC.

References

1. Xie L, Bourne P (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc Natl Acad Sci U S A 105: 5441–6.
2. Xie L, Xie L, Bourne P (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. Bioinformatics 25: i305–12.
3. Russell R (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 279: 1211–27.
4. Chen B, Fofanov V, Bryant D, Dodson B, Kristensen D, et al. (2007) The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs. J Comp Biol 14: 791–816.
5. Bryant D, Moll M, Chen B, Fofanov V, Kavraki L (2010) Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction. BMC Bioinformatics 11: 242. doi: 10.1186/1471-2105-11-242.
6. Binkowski T, Adamian L, Liang J (2003) Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns. J Mol Biol 332: 505–526.
7. Binkowski T, Joachimiak A (2008) Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. BMC Struct Biol 8: 45.
8. Dundas J, Adamian L, Liang J (2011) Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and nad binding proteins. J Mol Biol 406: 713–729.
9. Chen B, Bryant D, Fofanov V, Kristensen D, Cruess A, et al. (2007) Cavity scaling: automated refinement of cavity-aware motifs in protein function prediction. J Bioinform Comput Biol 5: 353–82.
10. Nussinov R, Wolfson H (1991) Efficient detection of three-dimensional structural motifs in biological macro-molecules by computer vision techniques. Proc Natl Acad Sci U S A 88: 10495–9.
11. Orengo C, Taylor W (1996) SSAP: Sequential Structure Alignment Program for Protein Structure Comparison. Methods Enzymol 266: 617–635.
12. Petrey D, Honig B (2003) GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. Methods Enzymol 374: 492–509.
13. Shindyalov I, Bourne P (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11: 739–47.
14. Yang AS, Honig B (2000) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol 301: 679–89.
15. Holm L, Sander C (1996) Mapping the protein universe. Science 273: 595–603.
16. Poirrette A, Artymiuk P, Rice D, Willett P (1997) Comparison of protein surfaces using a genetic algorithm. J Comput Aided Mol Des 11: 557–69.
17. Gibrat J, Madej T, Bryant S (1996) Surprising similarities in structure comparison. Curr Opin Struct Biol 6: 377–85.
18. Ye Y, Godzik A (2005) Multiple flexible structure alignment using partial order graphs. Bioinformatics 21: 2362–2369.
19. Rosen M, Lin S, Wolfson H, Nussinov R (1998) Molecular shape comparisons in searches for active sites and functional similarity. Protein Eng 11: 263–77.
20. Kinoshita K, Nakamura H (2003) Protein informatics towards function identification. Curr Opin Struct Biol 13: 396–400.
21. Kinoshita K, Nakamura H (2005) Identification of the ligand binding sites on the molecular surface of proteins. Protein Sci 14: 711–718.
22. Sael L, La D, Li B, Rustamov R, Kihara D (2008) Rapid comparison of properties on protein surface. Proteins 73: 1–10.
23. Chen B, Fofanov V, Kristensen D, Kimmel M, Lichtarge O, et al. (2005) Algorithms for structural comparison and statistical analysis of 3D protein motifs. In: Pac Symp Biocomput. volume 345, pp. 334–45.
24. Glaser F, Pupko T, Paz I, Bell R, Bechor-Shental D, et al. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19: 163–4.
25. Glaser F, Morris R, Najmanovich R, Laskowski R, Thornton J (2006) A method for localizing ligand binding pockets in protein structures. Proteins 62: 479–88.
26. Shatsky M, Shulman-Peleg A, Nussinov R, Wolfson H (2006) The multiple common point set problem and its application to molecule binding pattern detection. J Comput Biol 13: 407–28.
27. Blomberg N, Gabdoulline R, Nilges M, Wade R (1999) Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. Proteins 37: 379–387.
28. Chothia C, Lesk A (1986) The relation between the divergence of sequence and structure in proteins. EMBO J5: 823.
29. Martí-Renom M, Stuart A, Fiser A, Sánchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291–325.
30. Kolodny R, Petrey D, Honig B (2006) Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Curr Opin Struct Biol 16: 393–8.
31. Chen B, Bandyopadhyay S (2011) VASP-S: A Volumetric Analysis and Statistical Model for Predicting Steric Influences on Protein-Ligand Binding Specificity. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2011: 22–9.
32. Chen B, Bandyopadhyay S (2011) A Statistical Model of Overlapping Volume in Ligand Binding Cavities. In: Proceedings of the Computational Structural Bioinformatics Workshop (CSBW 2011). pp. 424–31.
33. Chen B, Bandyopadhyay S (2012) Modeling regionalized volumetric differences in protein-ligand binding cavities. Proteome Sci 10: S6.
34. Chen B, Bandyopadhyay S (2012) A regionalizable statistical model of intersecting regions in protein–ligand binding cavities. J Bioinform Comput Biol 10: 1242004. doi: 10.1142/S0219720012420048.
35. Chen B, Paul D (2012) A volumetric method for representing and comparing regions of electrostatic focusing in molecular structure. In: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM, pp. 242–249.
36. Godshall B, Chen B (2012) Improving accuracy in binding site comparison with homology modeling. Proceedings (IEEE Int Conf Bioinformatics Biomed): 662–669.
37. Chen B, Honig B (2010) VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity. PLoS Comput Biol 6: 11.
38. Record M, Anderson C, Lohman T (1978) 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. Q Rev Biophys 11: 103–178.
39. Warshel A, Russell S (1984) Calculations of electrostatic interactions in biological systems and in solutions. Q Rev Biophys 17: 283–422.
40. Matthew J (1985) Electrostatic effects in proteins. Annu Rev Biophys Biophys Chem 14: 387–417.
41. Honig B, Hubbell W, Flewelling R (1986) Electrostatic Interactions in Membranes and Proteins. Annu Rev Biophys Biophys Chem 15: 163–193.
42. Rogers N (1986) The modelling of electrostatic interactions in the function of globular proteins. Prog Biophys Mol Biol 48: 37–66.
43. Harvey S (1989) Treatment of Electrostatic Effects in Marcomolecular Modeling. Proteins 5: 78–92.
44. McLaughlin S (1989) The electrostatic properties of membranes. Annu Rev Biophys Biophys Chem 18: 113–36.
45. Sharp K, Honig B (1990) Electrostatic Interactions in Macromolecules: Theory and Applications. Annu Rev Biophys Biophys Chem 19: 301–332.
46. Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. Science 268: 1144–1149.
47. Nakamura H (1996) Roles of electrostatic interaction in proteins. Q Rev Biophys 29: 1–90.
48. Polticelli F, Honig B, Ascenzi P, Bolognesi M (1999) Structural determinants of trypsin affinity and specificity for cationic inhibitors. Protein Sci 8: 2621–2629.
49. Hendsch Z, Tidor B (1994) Do salt bridges stabilize proteins? A continuum electrostatic analysis. Protein Sci 3: 211–26.
50. Sindelar C, Hendsch Z, Tidor B (1998) Effects of salt bridges on protein structure and design. Protein Sci 7: 1898–1914.
51. O'Shea E, Rutkowski R, Kim P (1992) Mechanism of specificity in the Fos-Jun oncoprotein heterodimer. Cell 68: 699–708.
52. Kangas E, Tidor B (2001) Electrostatic Complementarity at Ligand Binding Sites: Application to Chorismate Mutase. J Phys Chem B 105: 880–888.
53. Lee L, Tidor B (1997) Optimization of electrostatic binding free energy. J Chem Phys 106: 8681–8690.
54. Voelcker H, Requicha A (1977) Geometric Modeling of Mechanical Parts and Processes. Comput J 10: 48–57.
55. Ju T, Losasso F, Schaefer S, Warren J (2002) Dual contouring of hermite data. ACM Trans Graph 21: 339–346.
56. McCoy A, Chandana E, Colman P (1997) Electrostatic complementarity at protein/protein interfaces. J Mol Biol 268: 570–584.
57. Zhang X, Bajaj C, Kwon B, Dolinsky T, Nielsen J, et al. (2006) Application of new multi-resolution methods for the comparison of biomolecular electrostatic properties in the absense of global structural similarity. Multiscale Model Simul 5: 1196–1213.
58. Elcock A (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 312: 885–96.
59. Murray D, Honig B (2002) Electrostatic control of the membrane targeting of c2 domains. Mol Cell 9: 145–154.
60. Gilson M, Honig B (1987) Calculation of electrostatic potentials in an enzyme active site. Nature 330: 84–86.
61. Gilson M, Sharp K, Honig B (1988) Calculating the electrostatic potential of molecules in solution: Method and error assessment. J Comput Chem 9: 327–335.
62. Menke M, Berger B, Cowen L (2008) Matt: local flexibility aids protein multiple structure alignment. PLoS Comp Biol 4: e10.
63. Laskowski R (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13: 323–30, 307–8.
64. Armon A, Graur D, Ben-tal N, Aviv R (2001) ConSurf: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Phylogenetic Information. J Mol Biol 307: 447–463.

65. Nayal M, Honig B (2006) On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites. Proteins 63: 892–906.
66. Dundas J, Ouyang Z, Tseng J, Binkowski T, Turpaz Y, et al. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34: W116–8.
67. Coleman R, Sharp K (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. J Mol Biol 362: 441–58.
68. Rocchia W, Alexov E, Honig B (2001) Extending the applicability of the nonlinear poisson-boltzmann equation: Multiple dielectric constants and multivalent ions. J Phys Chem B 105: 6507–6514.
69. Lorensen W, Cline H (1987) 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, pp. 163–170.
70. Nicholls A, Sharp K, Honig B (1991) Protein folding and association: insights from the interfacial and thermo-dynamic properties of hydrocarbons. Proteins 11: 281–96.
71. Connolly M (1983) Solvent-accessible surfaces of proteins and nucleic acids. Science 221: 709–713.
72. Schaer J, Stone M (1991) Face traverses and a volume algorithm for polyhedra. Lect Notes Comput Sc 555/1991: 290–297.
73. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–42.
74. Chen V, Arendall W, Headd J, Keedy D, Immormino R, et al. (2010) Molprobity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr D66: 12–21.
75. Keskin O, Ma B, Nussinov R (2005) Hot Regions in Protein Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues. J Mol Biol 345: 1281–1294.
76. Klapper I, Kagstrom R, Fine R, Sharp K, Honig B (1986) Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. Proteins 1: 47–59.
77. Rohs R, West S, Sosinsky A, Liu P, Mann R, et al. (2009) The role of dna shape in protein–dna recognition. Nature 461: 1248–1253.
78. Costantini S, Colonna G, Facchiano A (2008) Esbri: a web server for evaluating salt bridges in proteins. Bioinformation 3: 137.
79. Felsenstein J (1989) Phylip - phylogeny inference package (version 3.2): 164–166.
80. Yang AS, Honig B (2000) 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: 665–78.
81. Sievers F, Wilm A, Dineen D, Gibson T, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol 7: 539. doi: 10.1038/msb.2011.75.
82. Junier T, Zdobnov E (2010) The newick utilities: high-throughput phylogenetic tree processing in the unix shell. Bioinformatics 26: 1669–1670.
83. Schechter I, Berger A (1967) On the size of the active site in proteases. I. Papain. Biochem Biophys Res Commun 27: 157–162.
84. Gráf L, Jancsó a, Szilágyi L, Hegyi G, Pintér K, et al. (1988) Electrostatic complementarity within the substrate-binding pocket of trypsin. Proc Natl Acad Sci U S A 85: 4961–5.
85. Morihara K, Tsuzuki H (1969) Comparison of the specificities of various serine proteinases from microorganisms. Arch Biochem Biophys 129: 620–634.
86. Steitz T, Hendekson R, Blow D (1969) Structure of crystalline a-chymotrypsin 3. Crystallographic studies of substrates and inhibitors bound to the active site of a-chymotrypsin. J Mol Biol 46: 337–340.
87. Halangk W, Lerch M, Brandt-Nedelev B, Roth W, Ruthenbuerger M, et al. (2000) Role of cathepsin b in intracellular trypsinogen activation and the onset of acute pancreatitis. J Clin Invest 106: 773–781.
88. Sloane B, Moin K, Krepela E, Rozhin J (1990) Cathepsin b and its endogenous inhibitors: the role in tumor malignancy. Cancer Metastasis Rev 9: 333–352.
89. Khouri H, Plouffe C, Hasnain S, Hirama T, Storer A, et al. (1991) A model to explain the ph-dependentspecificity of cathepsin b-catalysed hydrolyses. Biochem J 275: 751–757.
90. Hasnain S, Hirama T, Huber C, Mason P, Mort J (1993) Characterization of cathepsin b specificity by site-directed mutagenesis. importance of glu245 in the s2-p2 specificity for arginine and its role in transition state stabilization. J Biol Chem 268: 235–240.
91. Storer A, Ménard R (1996) Recent insights into cysteine protease specificity: Lessons for drug design. Perspect Drug Discovery Des 6: 33–46.
92. Wieczerzak E, Rodziewicz-Motowid lo S, Jankowska E, Giełdón A, Ciarkowski J (2007) An enormously active and selective azapeptide inhibitor of cathepsin b. J Pept Sci 13: 536–543.
93. Musil D, Zucic D, Turk D, Engh R, Mayr I, et al. (1991) The refined 2.15 a x-ray crystal structure of human liver cathepsin b: the structural basis for its specificity. EMBO J 10: 2321.
94. Jones D, Bycroft M, Lubienski M, Fersht A (1993) Identification of the barstar binding site of barnase by nmr spectroscopy and hydrogen-deuterium exchange. FEBS Lett 331: 165–172.
95. Buckle A, Schreiber G, Fersht A (1994) Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0a resolution. Biochemistry 33: 8878–8889.
96. Schreiber G, Fersht A (1996) Rapid, Electrostatically Assisted association of proteins. Nat Struct Mol Biol 3: 427–431.
97. Meiering E, Bycroft M, Fersht A (1991) Characterization of phosphate binding in the active site of barnase by site-directed mutagenesis and nmr. Biochemistry 30: 11348–11356.
98. Schreiber G, Fersht A (1993) Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering. Biochemistry 32: 5145–5150.
99. Schreiber G, Fersht A (1995) Energetics of protein-protein interactions: Analysis ofthe barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol 248: 478–486.
100. Tissot A, Vuilleumier S, Fersht A (1996) Importance of two buried salt bridges in the stability and folding pathway of barnase. Biochemistry 35: 6786–6794.
101. Hartley R (1993) Directed mutagenesis and barnase-barstar recognition. Biochemistry 32: 5978–5984.
102. Cox A, Der C (2003) The dark side of ras: regulation of apoptosis. Oncogene 22: 8999–9006.
103. Shields J, Pruitt K, McFall A, Shaub A, Der C (2000) Understanding ras: it aint overtil its over. Trends Cell Biol 10: 147–154.
104. Goodsell D (1999) The molecular perspective: the ras oncogene. Oncologist 4: 263–264.
105. Nassar N, Horn G, Herrmann C, Scherer A, McCormick F, et al. (1995) The 2.2 å crystal structure of the ras-binding domain of the serine/threonine kinase c-raf1 in complex with rapla and a gtp analogue. Nature 375: 554–560.
106. Kyriakis J, App H, Zhang XF, Banerjee P, Brautigan D, et al. (1992) Raf-1 activates map kinase-kinase. Nature 358: 417–421.
107. Kiel C, Serrano L, Herrmann C (2004) A detailed thermodynamic analysis of ras/effector complex interfaces. J Mol Biol 340: 1039–1058.
108. Jaitner B, Becker J, Linnemann T, Herrmann C, Wittinghofer A, et al. (1997) Discrimination of amino acids mediating ras binding from noninteracting residues affecting raf activation by double mutant analysis. J Biol Chem 272: 29927–29933.
109. Herrmann C, Martin G, Wittinghofer A (1995) Quantitative analysis of the complex between p21 and the ras-binding domain of the human raf-1 protein kinase. J Biol Chem 270: 2901–2905.
110. Nassar N, Horn G, Herrmann C, Block C, Janknecht R, et al. (1996) Ras/rap effector specificity determined by charge reversal. Nat Struct Mol Biol 3: 723–729.
111. Akasaka K, Tamada M, Wang F, Kariya KI, Shima F, et al. (1996) Differential structural requirements for interaction of ras protein with its distinct downstream effectors. J Biol Chem 271: 5353–5360.
112. Gasper R, Thomas C, Ahmadian M, Wittinghofer A (2008) The role of the conserved switch ii glutamate in guanine nucleotide exchange factor-mediated nucleotide exchange of gtp-binding proteins. J Mol Biol 379: 51–63.
113. DeClue J, Stone J, Blanchard R, Papageorge A, Martin P, et al. (1991) A ras effector domain mutant which is temperature sensitive for cellular transformation: interactions with gtpase-activating protein and nf-1. Mol Cell Biol 11: 3132–3138.