doi:10.1016/j.jmb.2003.08.019 *J. Mol. Biol.* (2003) **332**, 1095–1113

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Identification of Substrate Binding Sites in Enzymes by Computational Solvent Mapping

Michael Silberstein[1], Sheldon Dennis[2], Lawrence Brown III[2] Tamas Kortvelyesi[2,3], Karl Clodfelter[1] and Sandor Vajda[2]*

[1]*Program in Bioinformatics Boston University, Boston, MA 02215, USA*

[2]*Department of Biomedical Engineering, Boston University 44 Cummington Street, Boston MA 02215, USA*

[3]*Department of Physical Chemistry, University of Szeged, 6701 Szeged, P.O. Box 105, Hungary*

**Corresponding author*

Enzyme structures determined in organic solvents show that most organic molecules cluster in the active site, delineating the binding pocket. We have developed algorithms to perform solvent mapping computationally, rather than experimentally, by placing molecular probes (small molecules or functional groups) on a protein surface, and finding the regions with the most favorable binding free energy. The method then finds the consensus site that binds the highest number of different probes. The probe–protein interactions at this site are compared to the intermolecular interactions seen in the known complexes of the enzyme with various ligands (substrate analogs, products, and inhibitors). We have mapped thermolysin, for which experimental mapping results are also available, and six further enzymes that have no experimental mapping data, but whose binding sites are well characterized. With the exception of haloalkane dehalogenase, which binds very small substrates in a narrow channel, the consensus site found by the mapping is always a major subsite of the substrate-binding site. Furthermore, the probes at this location form hydrogen bonds and non-bonded interactions with the same residues that interact with the specific ligands of the enzyme. Thus, once the structure of an enzyme is known, computational solvent mapping can provide detailed and reliable information on its substrate-binding site. Calculations on ligand-bound and apo structures of enzymes show that the mapping results are not very sensitive to moderate variations in the protein coordinates.

Keywords: structural genomics; X-ray structure; protein solvation; organic solvent; ligand binding

Introduction

A major challenge in structural genomics is the elucidation of biochemical and biological properties of enzymes, including the determination of amino acid residues that belong to the ligand/substrate-binding site.[1,2] The two main sources of information on specific molecular interactions are the structures of the enzyme, co-crystallized with various ligands (substrates, cofactors, inhibitors, products, and transition state analogs), and site-directed mutagenesis of the putative binding site residues. Since the available complexes provide complete structural characterization only for a fraction of the enzymes with known structure,[2] and mutational analyses are slow and labor-intensive, developing a method for determining the functional site on the basis of protein structure has been an important goal.[3]

A potentially useful strategy for determining ligand-binding sites on the surface of a protein is solvent mapping, i.e. solving the X-ray structure of the protein in a variety of organic solvents.[4,5]

Present addresses: S. Dennis, Biotechnology Research Institute, National Research Council of Canada, Montreal, Canada; L. Brown III, Fish and Neave, New York City, NY 10020,USA.

Abbreviations used: NMR, nuclear magnetic resonance; PDB, Protein Data Bank; CS-Map, computational solvent mapping; GRAMM, global range molecular matching; ACP, atomic contact potential; ACE, analytical continuum electrostatics; DMSO, dimethylsulfoxide; 2-PGA, 2-phosphoglycerate.

E-mail address of the corresponding author: vajda@bu.edu

Data for elastase[4–6] and thermolysin[7,8] show that the protein structure remains virtually identical to the native structure, and a limited number of organic molecules (typically 1 to 12) are associated with the protein surface in the first shell of water molecules. The power of the method arises from superimposing at least four or five structures of a protein solved in different solvents.[5] For enzymes, the probes cluster in the active site, forming a "consensus" site that delineates the binding pocket. All other binding sites are either in crystal contact, occur only at high ligand concentrations, or are in small, buried pockets that bind only a subset of the solvent molecules rather than all of them. The preferential binding of organic molecules to the active site has also been shown in aqueous solution by NMR methods.[9]

Since experimental solvent mapping requires repeated structure determination, and the protein may have to be cross-linked for added stability in organic solvents, the method is relatively expensive. We have developed algorithms to perform the mapping computationally by using small organic molecules as probes on the protein surface, and determining the consensus sites that bind a number of different probes.[10,11] The method has been first applied to hen egg-white lysozyme[9] and thermolysin,[7,8] because these proteins have been experimentally mapped using a number of organic solvents. In both cases, the probes cluster in the active site, in good agreement with the results of the mapping experiments.

While structural genomics efforts are likely to provide structures for an increasing number of poorly characterized proteins, few computational tools are available for determining the functionally important residues,[12–20] even when the structure is known, and hence computational mapping is potentially important. However, several problems need to be addressed before the method can be considered a useful tool. First, both experimental and computational mapping methods have been applied only to a handful of proteins. Thus, the general applicability of the approach is not at all clear, and more proteins should be mapped to see if the organic solvents cluster in the active site, irrespective of their size and polarity.[4–8] Second, assuming that such clustering occurs, it is still necessary to study a number of well-known enzymes, and to carefully evaluate the information provided by mapping.

Here we address the above problems. First, the mapping algorithm is applied to thermolysin, the protein with the most extensive experimental mapping data available. The results are compared to the ligand positions in the X-ray structures of the protein, determined in aqueous solutions of isopropanol, acetone, acetonitrile, and phenol.[7,8] The interactions between the probes and particular residues of the protein are also compared to the interactions seen in the known complexes of thermolysin with various ligands (substrate and transition state analogs, products, and inhibitors), extracted from the RCSB PDB. Second, we map six enzymes, enolase, fructose-1,6-biphosphatase, ribonuclease T_1, trypsin, haloalkane dehalogenase, and triosephosphate isomerase, that have no experimental mapping data, but whose binding sites are well characterized. These particular enzymes were selected because their substrate binding sites are not in the largest pockets,[21] thereby avoiding the possibility that the mapping finds the largest crevice on the protein surface, which is frequently the case when using simple geometric methods.[15–18] We map each protein using acetone, urea, dimethylsulfoxide (DMSO), isopropanol, *t*-butanol, and phenol as probes, identify the consensus site, and compare the results to the interactions extracted from all complexes of these enzymes with various ligands in the PDB. Mapping is performed both for the ligand-bound and the apo structures of each enzyme.

With the exception of haloalkane dehalogenase, which binds substrates that are smaller than some of the probes, the consensus site with the highest number of different probes occurs in a major subsite of the enzyme active site. Clusters at nearby locations indicate other subsites of the active site, and hence are also considered in the further analysis. As we will show, the selected clusters generally delineate the entire active site. In particular, the residues of an enzyme that most frequently interact with the probes also bind many ligands (substrate analogs, products, and inhibitors). Apart from particular cases in which either the binding site is not accessible to probes, or part of the protein is not present in the calculations, the mapping always finds most of the binding site residues, and there are very few false positives, i.e. residues hit by the probes that do not belong to the binding site. Our results suggest that the clustering of organic solvents in the binding sites of enzymes is a general property that applies to all enzymes, and thus solvent mapping is a potentially important tool to study poorly characterized enzymes if their structure is available.

Results

Identification of the thermolysin binding site

Thermolysin is zinc endopeptidase with two quasi-spherical domains separated by a large groove containing the active site zinc ion, coordinated by H142, H146, E166, and one water molecule.[22,23] Since structures are available for more than 20 complexes of thermolysin with bound ligands, including transition-state analogue inhibitors, the binding site and catalytic mechanism are well understood. English *et al.*[7,8] determined high resolution crystal structures of thermolysin, generated from crystals soaked in aqueous solutions of isopropanol, acetone, acetonitrile, and phenol. An increasing number of solvent interaction sites could be identified as the

solvent concentration was increased, up to 12 bound molecules in the case of isopropanol. However, the S_1' subsite was shown to be exceptional on two accounts. First, at low solvent concentrations, this is the only binding site, and the concentrations must be substantially increased (up to 80% in the case of isopropanol) before binding occurs at any other location. Second, superimposing all structures shows that S_1' is the only site where all four solvent molecules bind. Isopropanol also binds to subsites S_1 and S_2, but only at high concentrations.[7] Figure 1a shows the ligand positions in the active site of superimposed thermolysin structures solved in 10% isopropanol,[7] 50% acetone, 50% acetonitrile, and 50 mM phenol.[8]

Computational mapping was applied to a thermolysin structure (RCSB PDB code: 2tlx), co-crystallized with the dipeptide Val-Lys, a cleavage product, which is also shown in Figure 1 for reference. The PDB does not include any thermolysin structure without a bound ligand. We have removed the peptide, the active site Zn^{2+}, and all crystallographic water molecules, and mapped the entire protein surface using both the CS-Map algorithm and the GRAMM-based approach to mapping. In agreement with the experimental data,[7,8]

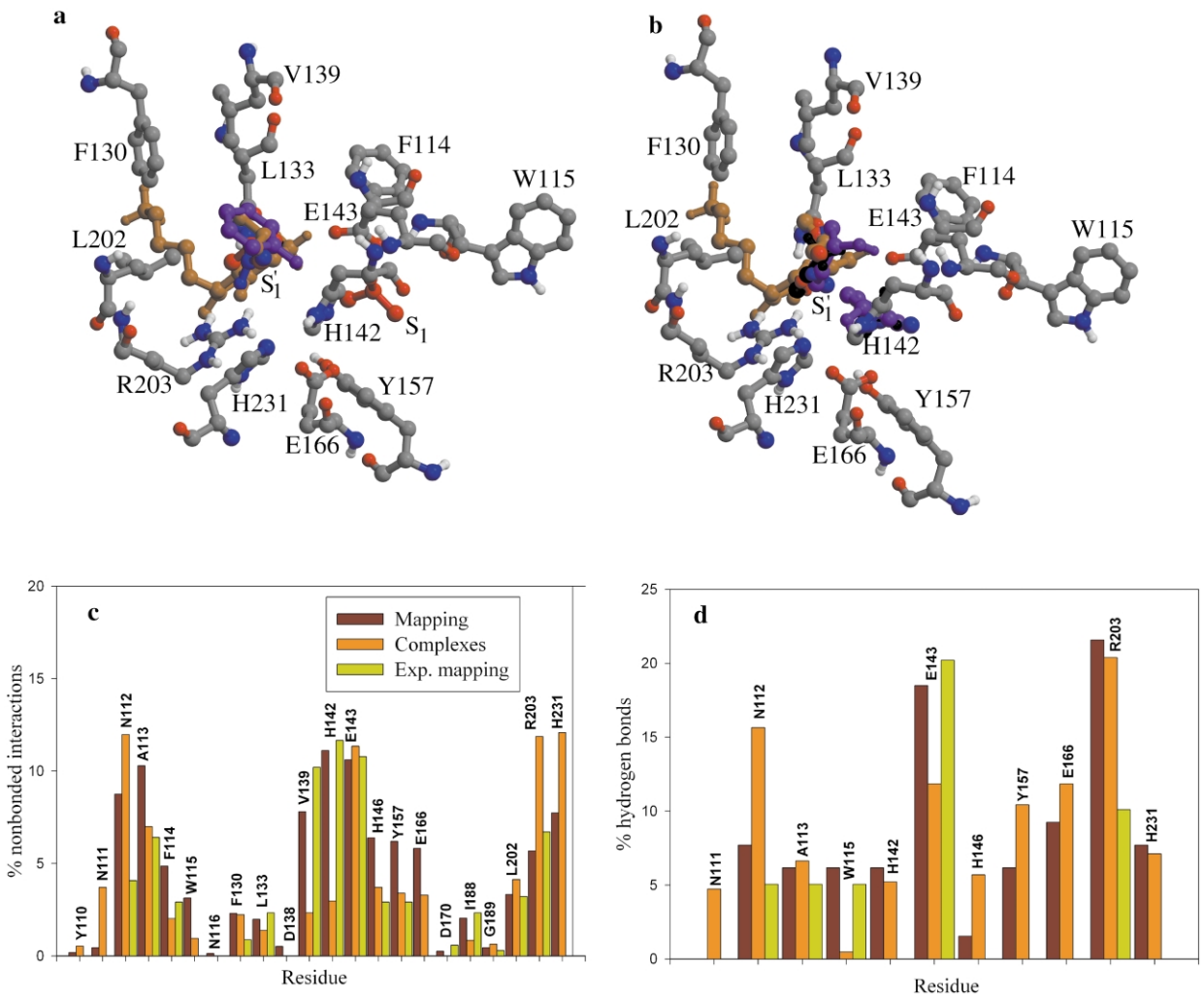

Figure 1. Experimental and computational mapping of thermolysin. a, Superimposed ligand positions in thermolysin co-crystallized with the V-K dipeptide (2tlx), and in structures solved in 10% isopropanol,[7] 50% acetone, 50% acetonitrile, and 50 mM phenol.[8] The color scheme used for the ligands is ochre, V-K dipeptide; red, isopropanol; blue, acetone; black, acetonitrile; and purple, phenol. For the protein side-chains we use the standard atomic colors, i.e.carbon, grey; oxygen, red; nitrogen, blue; and hydrogen, white. All solvents bind in the S_1' pocket, and isopropanol also binds at the S_1 and S_2 sites (the latter is not shown). b, Computational mapping of thermolysin using the CS-Map algorithm. The Figure shows the main consensus site in the S_1' pocket that binds all four solvents (Table 1, site 1), and the second consensus site close to S_1, which binds three solvents (Table 1, site 2). c, Distribution of intermolecular non-bonded interactions among thermolysin residues. The interactions were determined from three sources: computational mapping; extracted from 43 complexes of thermolysin with different ligands in the RCSB PDB database; and experimental mapping.[7,8] Computational mapping results are based on the interactions found between various thermolysin residues and the probes in the main consensus site. The Figure shows the union of the three sets of interacting residues as determined by the three methods. d, The same as c, but for hydrogen bonds rather than non-bonded interactions.

isopropanol, acetone, acetonitrile, and phenol were used as probes. For each method, the five lowest free energy clusters for each probe were superimposed (see Methods) to find the consensus sites shown in Table 1. The integers in this Table represent the ranking of probe clusters, e.g. the consensus site 1 found by CS-Map is located in the S_1' pocket, contains the fourth lowest free energy cluster of acetone, the second lowest free energy cluster of phenol, and so on. For reference, Table 1 also shows the distances from the center of a isopropanol molecule, labeled as IPA1 in the X-ray structures of thermolysin solved in isopropanol,[7] which binds in the S_1' pocket.

According to Table 1, CS-Map finds a single consensus site that binds all four molecules in the S_1' pocket (Figure 1b), in good agreement with the experimental data. Phenol, acetone, and acetonitrile cluster at two additional positions, the first being close to the S_1 subsite (Table 1, consensus site 2). Although the computational mapping also places the lowest free energy isopropanol cluster into the S_1 pocket, this location is distinct from the second consensus site that includes the other three solvents. GRAMM also finds a single consensus site, with four solvents bound, in the S_1' site, even closer to IPA1 than the one found by CS-Map. The method places the lowest free energy clusters of acetone and isopropanol, and the second lowest free energy clusters of acetonitrile in an almost completely buried pocket (Table 1, site 2 for GRAMM). It is interesting that this pocket has been shown experimentally to bind three of the four solvents at elevated concentrations, but the exception was acetonitrile rather than phenol.[8] While CS-Map and GRAMM yield the same consensus site for thermolysin, the results also show that GRAMM has a higher tendency to put probes in largely buried pockets. We note that combining the results from CS-Map and GRAMM we find the main consensus site to include eight probe clusters, whereas at most three clusters overlap at any other location. As will be discussed further, such combination of results from the two methods generally helps to better discriminate the consensus site from other locations that bind some of the probes.

In the case of thermolysin we have used the CS-Map results to characterize the binding site. Since the main consensus site (Table 1, site 1) and site 2 are within 3 Å to each other, both were considered in the analysis (see Methods), and the clusters at both locations were divided into subclusters. Sub-clustering shows that each probe molecule binds in a number of rotational states, with the non-polar moiety located in a hydrophobic pocket defined by the side-chains of L202, F130, L133, and F114 (Figure 1b), and the polar part pointing toward various polar patches on the protein, in some cases forming one or two hydrogen bonds.[10] The X-ray structures also suggest that, apart from the solvent in the buried pocket, the bound molecules are fairly mobile, generally with *B* factors around 60, and the existence of several possible binding modes has been noted by the crystallographer.[7,8] Selecting a representative conformation from each sub-cluster, we counted the non-bonded interactions and hydrogen bonds between the probes and the protein, and determined their distribution among the amino acid residues.

The above distributions were compared to the ones based on experimental solvent mapping,[7,8] as well as to the interactions extracted from the 23 thermolysin complexes in the RCSB PDB. The binding of substrate and transition state analogs, products, and inhibitors always involves the largely hydrophobic S_1' sub-site, with the substrates and longer inhibitors extending toward sub-sites S_1 and S_2. At least four hydrogen bonds are formed in each complex, most frequently with the side-chains of R203, E143, Y157, and N112, and with the polar backbone atoms of Y115, A113, and N111. As shown in Figure 1c and d, the residues that are important for the binding of specific ligands also interact with many probes.

While experimental solvent mapping identifies only a subset of the important residues (E143, R203, N112, and A113), computational mapping provides essentially complete information on the residues in the binding site, in terms of both non-bonded interactions and hydrogen bonds (Figure 1c and d, respectively). Indeed, H142, H146, and E166 coordinate the Zn^{2+} in the active site, E143

Table 1. Ranking of probe clusters within the consensus sites for thermolysin

Algorithm	Consensus site[a]	Probe			
		Acetone	Phenol	Isopropanol	Acetonitrile
CS-Map	**1 (S_1')**	**4 (0.8)**	**2 (0.7)**	**2 (0.7)**	**3 (0.3)**
	2 (S_1)	2 (3.7)	1 (3.5)	–	1 (3.9)
	3	1 (18.1)	4 (17.3)	–	2 (17.4)
GRAMM	**1 (S_1')**	**5 (0.2)**	**3 (0.3)**	**4 (0.7)**	**3 (0.1)**
	2	1 (14.5)	–	1 (14.4)	2 (14.2)
	3	2 (20.1)	–	2 (20.2)	1 (19.8)
	4	3 (17.5)	–	3 (16.9)	5 (16.7)

Distance of each cluster center from the isopropanol position IPA1[7] is shown in parentheses.
[a] The two consensus sites in the substrate binding region are shown in bold.

serves as the general base, Y157 and H231 provide further stabilization of the transition state, while N112 and the backbone of A113 form hydrogen bonds with the leaving group.[22,23] While the same residues interact with many of the probes, the mapping results reflect the importance of each residue for substrate binding rather than for catalytic activity. For example, the mapping finds the highest number of hydrogen bonds for R203 (Figure 1d) which does not directly participate in hydrolysis, but forms hydrogen bonds with the carbonyl group of a residue at the P_1' position, and is known to be crucial for substrate binding.[24] The mapping does not find D226, which is part of the catalytic mechanism,[22] but the D226A mutation introduces only a minor perturbation in the activity.[25] It is important that we do not find any false positives, i.e. amino acid residues hit by the probes that are not part of the binding site.

Mapping and binding site identification in model enzymes

In our previous paper[10] we mapped hen egg-white lysozyme, which binds polysaccharides and has a very large cleft six saccharide units long. The mapping placed the lowest free energy clusters for each of eight different solvents in sub-site C of the binding site, in good agreement with intermolecular Overhauser effects that show site C to bind, almost exclusively, all eight compounds in aqueous solution.[9] Since finding a large binding site was really easy, here we study six enzymes, enolase (1ebg), ribonuclease T_1 (1rnt), triosephosphate isomerase (2ypi), fructose-1,6-biphosphatase (1fbc), trypsin (1tng), and haloalkane dehalogenase (2dhc), that all have relatively small binding sites.[21] This eliminates the possibility that the mapping simply finds the largest pocket on the protein surface. All ligands, including ions and crystallographic water molecules, were removed before the mapping by the less reliable, but faster GRAMM-based approach. Acetone, urea, DMSO, isopropanol, *t*-butanol, and phenol were used as probes. The five lowest free energy clusters of each probe were superimposed (see Methods) to find the consensus sites shown in Table 2.

Enolase (1ebg) catalyzes the dehydration of 2-phospho-D-glycerate (2-PGA) during glycolysis, and contains two Mg ions in the binding site.[26–28] Computational mapping, applied to the protein finds only one position at which clusters of all six probes overlap (Table 2, site 1 and Figure 2a). As shown in Table 2, these clusters are the lowest free energy ones for all solvents but isopropanol, which has its second lowest free energy cluster at this location. The consensus clustering occurs in the active site, with cluster centers less than 1 Å away from the position of 2-PGA. All other consensus sites are formed by the clusters of four or less different probes (Table 2).

In the clusters belonging to the main consensus site, the probes interact with all the residues that participate in the catalytic mechanism (Figures 3a and 4a). These residues are K345, the catalytic base; E211, the catalytic acid which donates proton to C3 hydroxyl group of 2-PGA; K396, interacting with one of the C1 carboxylate oxygen atoms to stabilize the redistribution of negative charge formed in the intermediate; E168, which may be essential for proper orientation of K396 and E211, as well as pK_a adjustment of these residues; and H373 which interacts with C3 hydroxyl oxygen atom of 2-PGA.[26–28] However, the highest numbers of contacts are found for residues that are important for substrate binding rather than catalysis. These are S39, which coordinates the lower affinity Mg ion,[29] and H159, R374, and S375 that interact with the phosphate group of the substrate (Figure 3a).[26–28] Recently, an alternative mechanism has been proposed in which H159 serves as the catalytic base,[30] but later it was shown that the H159A mutant of yeast enolase still has 0.2% of the native activity.[31]

In general, there is an excellent agreement between the non-bonded interactions predicted by the mapping, and those seen in the complexes of enolase with various ligands (Figure 3a). According to Figure 4a, the predicted and observed hydrogen bonds differ substantially more. The main deviation is due to S39, which coordinates the lower affinity Mg^{2+} through its backbone carbonyl and side-chain hydroxyl. Mapping has been performed without metal ions, and S39 hydrogen bonds with almost all probes. By contrast, the side-chains of D246, E295, and D320 that coordinate the higher affinity Mg^{2+}, did not show a strong tendency to form hydrogen bonds (Figure 4a), although they interact with many probes (Figure 3a). All other hydrogen-bonding groups are correctly identified by the mapping.

Ribonuclease T_1 (1rnt) is an extensively studied enzyme that catalyzes the hydrolysis of RNA at guanylyl residues.[32] According to the mapping, we have all six solvents clustering in the active site (Table 2, site 1). The probes interact with the catalytic residues E58, H92, and H40,[32] but the residues with the highest number of contacts and hydrogen bonds are in the guanine-binding loop Y42–E46 (Figures 3b and 4b). These latter residues render ribonuclease T_1 guanine-specific through a series of intermolecular hydrogen bonds, and their mutations affect the dissociation constant of the enzyme–substrate complex but do not affect the turnover rate.[33–35] We find the strongest hydrogen bonding affinity for N43, and large numbers of non-bonded interactions for Y45 and Y42. The latter is known to contribute significantly to guanine binding through a face-to-face parallel stacking, and the interactions with the side-chain are revealed by the mapping (Figure 3b). By contrast, for N43 and Y45 most interactions are with the backbone, i.e. the NH group of N43 interacts with guanine N7, and the NH of Y45 hydrogen bonds to O6 of guanidine moiety.[33] The Y45 side-chain is also important, and this

Table 2. Free energy ranking of probe clusters, mapped by GRAMM, within the consensus sites for the bound conformations of six enzymes

		Probe					
Protein	Con. site[a]	Acetone	Urea	DMSO	Isopropanol	*t*-Butanol	Phenol
1ebg	**1**	**1 (0.26)**	**1 (0.27)**	**1 (0.48)**	**2 (0.81)**	**1 (0.39)**	**1 (0.36)**
	2	–	3 (10.11)	2 (9.83)	3 (9.14)	2 (9.10)	2 (11.10)
	3	4 (20.61)	–	4 (20.99)	–	4 (20.83)	4 (20.59)
	4	2 (14.38)	4 (14.38)	3 (14.58)	–	–	5 (15.26)
	5	3 (25.30)	–	–	5 (24.52)	–	3 (24.23)
	6	5 (9.00)	–	5 (8.08)	4 (7.44)	–	–
1rnt	**1**	**2 (0.51)**	**4 (0.35)**	**1 (0.37), 4 (0.75)**	**5 (0.48)**	**4 (0.56)**	**1 (0.21)**
	2	3 (15.93)	5 (14.21)	5 (15.04)	–	5 (15.34)	5 (14.28)
	3	–	1 (10.61)	2 (11.76)	4 (12.38)	–	2 (12.31)
	4	–	–	3 (10.24)	1 (12.02)	–	4 (11.65)
	5	–	3 (14.24)	–	2 (15.92)	–	3 (14.42)
	6	1 (11.33), 5 (12.50)					
2ypi	**1**	**1 (0.49)**	**2 (0.76)**	**1 (0.46)**	**1 (0.70)**	–	**1 (0.26)**
	2	2 (20.14)	3 (19.35)	3 (20.75)	4 (20.17)	–	–
	3	–	–	–	3 (6.74)	1 (5.71)	2 (5.52)
	4	–	–	2 (17.48)	–	–	4 (16.14)
	5	3 (12.79)	–	–	2 (12.53)	–	–
	6		5 (17.51)		5 (16.74)		
	7	5 (12.75)		4 (12.60)			
1tng	**1**	**1 (0.30)**	**4 (0.65)**	**1 (0.17)**	**1 (0.60)**	**1 (0.32)**	**1 (0.50)**
	2	2 (17.46)	3 (18.01)	2 (17.44)	2 (16.92)	–	4 (17.18)
	3	4 (15.30)	–	3 (14.91)	3 (15.27)	2 (15.89)	2 (15.72)
	4	–	–	4 (16.54)	5 (17.61)	4 (17.85)	5 (15.94)
1fbc	**1**	**1 (1.41)**	**1 (0.43)**	**2 (0.59)**	**5 (0.80)**	**2 (0.61)**	**3 (0.92)**
	2	**5 (1.72)**	**3 (1.27)**	**4 (0.44)**	**1 (1.18)**	**5 (1.50)**	**2 (0.87)**
	3			**3 (0.40)**			
	4	2 (11.86)	2 (10.06)	5 (11.76)	2 (9.91)	4 (11.33)	1 (9.64)
	5	–	–	1 (22.85)	–	–	5 (23.46)
	6				4 (17.50)		4 (17.38)
2dhc	**1**	**1 (0.62)**	**1 (0.56)**	–	**1 (0.78)**	–	–
	2	3 (9.84)	3 (7.80)	3 (9.71)	3 (8.12)	1 (9.87)	2 (8.02)
	3	2 (10.81)	4 (10.32)	1 (10.23), 5 (10.36)	4 (9.80)	3 (10.49)	–
	4	4 (12.40)	2 (11.40)	4 (13.15)	5 (12.62)	–	–
	5	–	–	2 (12.23)	–	–	1 (11.37)
	6		5 (22.72)		2 (23.26)		
	7	5 (17.06)					5 (16.32)

Distance of each cluster center from the ligand in the X-ray structure of the enzyme is shown in parenthesis.
[a] Consensus sites in the substrate binding region are shown in bold.

residue, often referred to as the lid of the guanine-binding site, interacts with the highest number of probes.[32] Side-chain effects are also important for F100, which frequently interacts both with the probes in the mapping (Figure 3b) and with the ligands in the ribonuclease complexes, and is assumed to enhance the electrostatic interactions between substrate and active site *via* a change in the dielectric constant.[32] As expected, this residue forms many contacts with the probes (Figure 3b), but no hydrogen bonds (Figure 4b).

Triosephosphate isomerase (2ypi) catalyzes the tautomerization of dihydroxyacetone phosphate (DHAP), and is the first example of the TIM fold that occurs in variety of enzymes.[36] The largest consensus site found by the mapping is formed by lowest free energy clusters of acetone, DMSO, isopropanol, and phenol, and the second lowest free energy cluster of urea (Table 2, site 1 and Figure 2c). All other consensus sites have four or fewer different probes. At the main consensus site, the probes interact with N10, K12, H95, and E165 that are directly involved in catalysis (Figure 3c), the latter serving as the base.[36] The remaining nonbonded interactions are with three separate groups of residues on loops 6 (residues 166 to 176), 7 (residues 209 to 214), and 8 (residues 230 to 240), and bind to the phosphate group to the substrate.[36–38] In particular, the NH groups of G171, G210, S211, G232, and G233 make hydrogen bonds with the phosphate oxygen atoms. The hydrophobic side-chains of I170, L230, and V231 are also very important for binding (Figure 3c), which may explain the high sequence conservation at these positions.[38]

For trypsin (1tng) the only consensus site with six solvents is in the S_1 pocket (Table 2, site 1; see Figure 2d), formed by the lowest free energy clusters of all probes but urea, which has its fourth lowest free energy cluster at this location. The

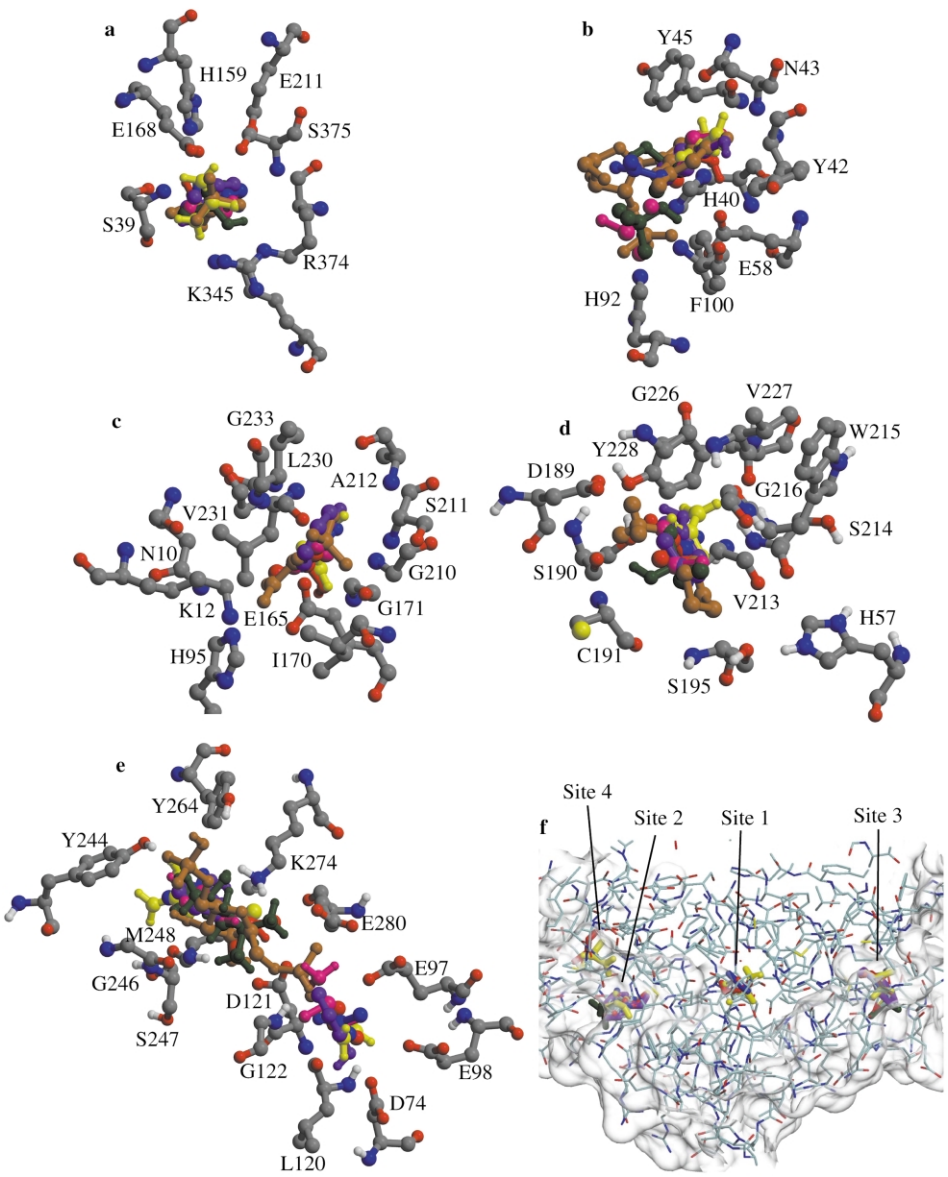

Figure 2. Consensus sites found by computational mapping, superimposed with the specific ligand of the enzyme. The color scheme for the ligands is ochre, ligand in X-ray structure of the enzyme; blue, acetone; yellow, urea; pink, DMSO; red, isopropanol; green, *t*-butanol; and purple, phenol. a, Enolase (1ebg), ligand: phosphonoacetohydroxamate. b, Ribonuclease T_1 (1rnt), ligand: 2′-guanylic acid. c, Triosephosphate isomerase (2ypi), ligand: 2-phosphoglycolate. d, Trypsin (1tng) ligand: aminomethylcyclohexane. e, Fructose-1,6-bisphosphatase (1fbc) ligand: 2,5-anhydroglucitol-1,6-biphosphate. Sites 1 and 2 are shown overlapping the two ends of the ligand. f, Haloalkane dehalogenase (2dhc), ligand: 1,2-dichloroethane. Site 4 is the putative ligand collision site.[30] The mapping suggests that the substrate bound at site 4 is shifted toward site 2 (one of the two main consensus sites), and then enters the channel and moves toward the catalytic site (site 1). The role of the second main consensus site (site 3) is not clear.

probes establish interactions with the specificity determining residues D189 and G216, and with catalytic residues S195 and H57[39,40] (Figure 3d). Most contacts occur with S190 and the backbone atoms of C191, W215, S214, and V213. The last three residues are known to fix the scissile bond of the substrate in a fixed orientation.[39,41] The side-chain of V213 is at the bottom of the S_1 pocket, and interacts with the probes more often than it is seen in the complexes of trypsin with various ligands. Similarly, the residue G226, the backbone of V227, and the side-chain of Y228 are all deep in the binding site, and thus are more available to the probe than to the larger substrate, resulting in some overprediction of their interactions (Figure 3d). The mapping finds many hydrogen bonds with S190, formed both by the backbone and by the side-chain (Figure 4d). Since S190 hydrogen bonds with the P_1 side-chain,[41] this prediction is correct. The probes also form hydrogen bonds with W215 and C191 that contact the P_3 and P_1 residues, respectively, in the substrate, but rarely form

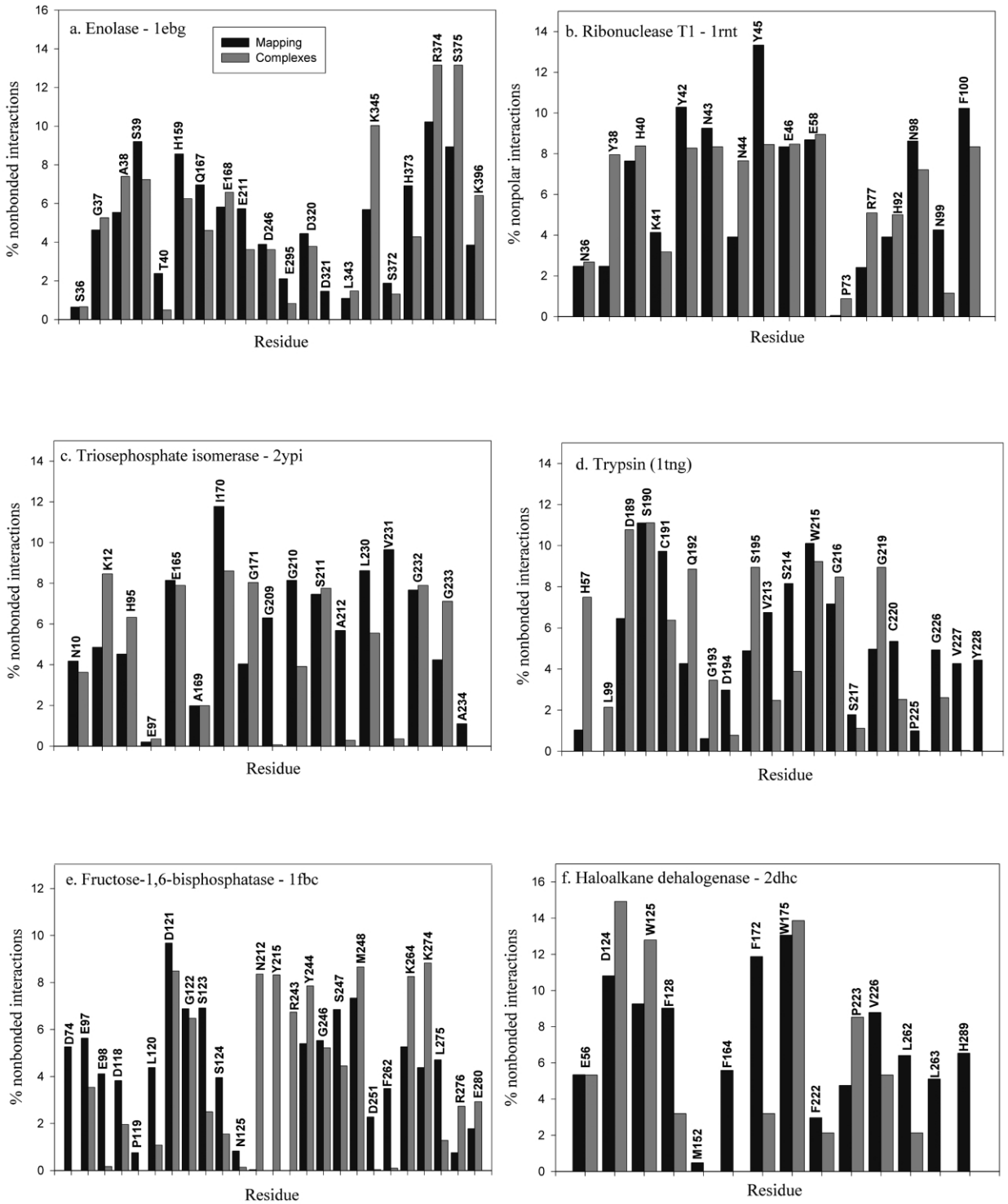

Figure 3. Distributions of intermolecular non-bonded interactions. The mapping results are based on the interactions found between the probes in the consensus site and the residues of the protein. The main consensus site is considered for all enzymes except haloalkane dehalogenase, where the probes in site 1 are used. The statistics on enzyme–ligand interactions are extracted from all enzyme–ligand complex structures in the RCSB PDB (see Methods). The Figure shows the union of the two sets of interacting residues as determined by the two methods, i.e. computational mapping and extraction from the complexes.

hydrogen bonds. The agreement is good for G219, which hydrogen bonds with the P_1 side-chain.

Fructose-1,6-bisphosphatase is a tetrameric enzyme that hydrolyzes fructose-1,6-bisphosphate in the presence of divalent cations.[42] The mapping of this protein yields three consensus sites with six solvents. Two of these sites are only 1 Å from each other (Table 2, sites 1 and 2), and together include

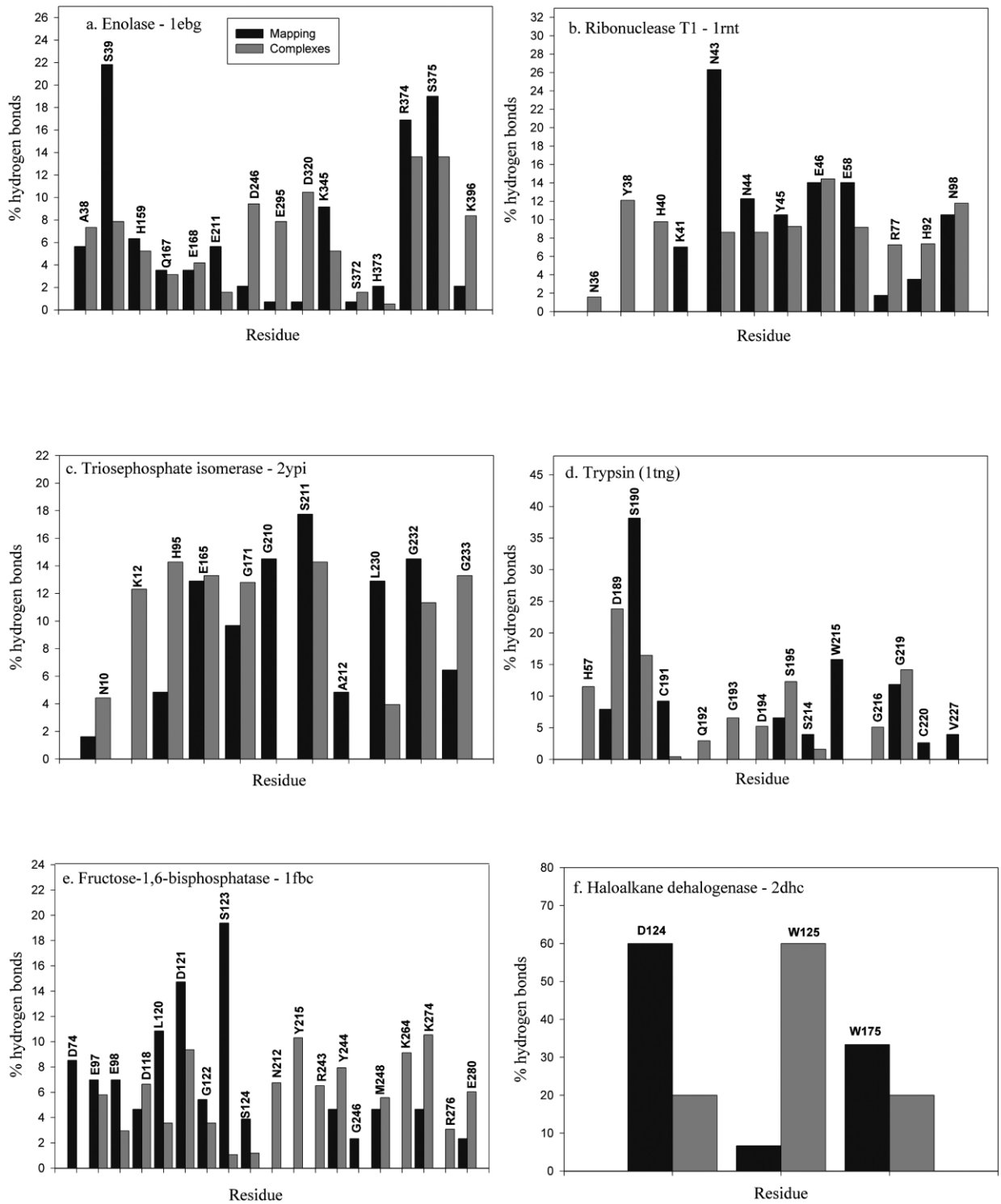

Figure 4. Distributions of intermolecular hydrogen bonds. The mapping results are based on the interactions found between the probes in the consensus site and the residues of the protein. The main consensus site is considered for all enzymes except haloalkane dehalogenase, where the probes in site 1 are used. The statistics on enzyme–ligand hydrogen bonds are extracted from all enzyme–ligand complex structures in the RCSB PDB (see Methods). The Figure shows the union of the two sets of interacting residues as determined by the two methods, i.e. computational mapping and extraction from the complexes.

the lowest free energy clusters for acetone, urea, and isopropanol, and the second lowest free energy clusters for DMSO, *t*-butanol, and phenol. As described in Methods, nearby consensus sites usually delineate different subsites of the active site, and this is really the case for this protein, site 1 being close to the location of the sugar ring in the enzyme–substrate complex, and site 2

surrounding the 2-phosphate group of the substrate (Figure 2e). Site 3 in Table 2 also binds six solvents, but with substantially higher free energies than sites 1 and 2 for four of the six probes.

The probes in the consensus site, formed by sites 1 and 2, interact with the catalytic bases D74 and E98,[42,43] but the most frequent interactions occur with residues that bind the ligand or one of the cations (Figure 3e). These are: L120, G122, S123, S124, and N125, which form hydrogen bonds with the 2-phosphate group of the substrate;[42] S247, M248, and K274, which contact the sugar ring;[41,42] Y244 and K264, which hydrogen bond to the 6-phosphate group;[43] and E97, D118, D121, R276, and E280, which coordinate with the cation.[42–44] The mapping does not find any interactions with N212, Y215, and R243. This result is due to mapping only a monomer of the dimeric protein. Since N212, Y215, and R243 are in a crevice formed by residues from two subunits, restricting consideration to a single subunit in the mapping removes the pocket, and the probes will not bind in this region.[42] In fact, R243 of one subunit interacts with the ligand bound to another subunit, and hence will not appear to be important for binding if only one subunit is considered.

As emphasized, mapping generally reflects the importance of residues for binding rather than for catalytic activity. Therefore, it is somewhat unexpected that it shows D74 and E98 to be relatively important (Figures 3e and 4e), in spite of the fact that these residues interact rarely with the ligand, but were proposed to act as the catalytic base, abstracting protons from the metal–hydroxide complex.[45] Apart from the minor differences we have mentioned, the agreement between predicted and observed non-bonded interactions is remarkably good (Figure 3e). The hydrogen-bonding residues are also correctly identified by the mapping, but the predicted frequency is less accurate than for the non-bonded interactions (Figure 4e). For example, S123 hydrogen bonds with the 2-phosphate group of the substrate[41] and hence is expected to interact with the probes, but the frequency of predicted hydrogen bonds is disproportionately large. The same applies to D121 that coordinates with a metal ion in the active site.[43,45]

For the haloalkane dehalogenase (2dhc) we find two consensus locations (Table 2, sites 2 and 3; see Figure 2e), both with six different solvents. Since neither of the two is close to the bound substrate, ethylene dichloride, it may appear that the mapping has failed, but this is not the case. Haloalkane dehalogenase binds very small ligands, such as ethylene dichloride, and the binding site is in the middle of a long and narrow channel.[46,47] Consensus site 1 (see Figure 2f and Table 2), which includes the lowest free energy clusters for acetone, urea, and isopropanol, is exactly at this location, and the correlation between the frequency of non-bonded interactions revealed by the mapping and seen in the complexes of haloalkane dehalogenase with ligands is very good (Figure 3e). However, the larger solvents are unable to enter the channel. The two large consensus sites (sites 2 and 3) are at the two ends of the channel. Thus, all ligands preferentially bind at these locations, but only the small ligands (acetone, urea, and isopropanol) can move into the channel toward the catalytic site (Table 2, site 1). An additional consensus site of three solvents (Table 2, site 4) is adjacent to site 3, and the existence of a collision complex formed during halide import is supported by both crystallographic and kinetic evidence, the latter involving the mutations of residues T197 and F294 that are at consensus site 4.[47] Taken together, these kinetic data and the mapping results strongly suggest that the substrate enters the channel at site 2 (Figure 2f). The role of binding at site 3 is not clear.

Mapping the apostructures of the model enzymes

In order to study the effect of conformational changes on the mapping results, the same GRAMM-based algorithm was applied to the apo structures of the six enzymes. Since more probes generally yield higher reliability of results, we have added acetonitrile as the seventh probe. Table 3 shows the derived consensus sites. For four of the apo structures we were still successful in locating the known substrate-binding site, with very little deviation from the analysis of the bound forms. For triosephosphate isomerase (1ypi), six out of the seven probes cluster at the experimental binding site, while the next largest consensus site has only four different solvents. Trypsin (3ptn) also has six probes clustered at the experimental S_1-binding site, while its closest runner-up site has five solvents clustered. For fructose-1,6-bisphosphatase (2fbp), both subsites of the binding site (as described earlier) were successfully found with seven and six solvent probes, respectfully. The next most occupied site had only four clustered probes. Finally, for haloalkane dehalogenase (2had), the most occupied consensus site consisted of one of the two ends of the channel that leads to the internal active site (Table 3, site 2). This site had all seven probes successfully mapped to it. In addition to this, the active site (site 1) had four solvents clustered, while the additional end of the channel (site 3) also had four probes. Therefore, as stated earlier, although the main consensus site for haloalkane dehalogenase is not the internal active site, the observation that most probes cluster at the entrance of the deep internal channel by which the substrate must traverse implies a successful mapping result.

The fast, GRAMM-based mapping algorithm was less successful in locating the binding site of the remaining two enzymes, enolase and ribonuclease T_1 when their apo structures were used. For enolase (1ebh) five different probes cluster at the binding site (Table 3, site 1), but a second location, about 16 Å away, also binds five probes

Table 3. Free energy ranking of probe clusters mapped by GRAMM for six apoenzymes

		Probe						
Protein	Con. site[a]	Acetone	Urea	DMSO	Isopropanol	*t*-Butanol	Phenol	Acetonitrile
1ebh	**1**		**5 (1.17)**	**3 (0.72)**		**4 (0.35)**	**1 (0.55)**	**1 (0.66)**
	2	1 (16.27)	4 (15.99)	2 (16.28)	1 (15.94)			2 (16.16)
	3	4 (23.13)		5 (23.02)	2 (22.18)			
	4			1 (7.82)		3 (7.38)		3 (8.13)
	5					1 (9.45)	2 (8.42), 3 (10.43)	
	6	2 (25.00)			4 (24.47)		4 (24.12)	
9rnt	1	5 (6.80)	4 (5.72)	4 (6.66)	5 (6.68)	4 (6.21)		3 (6.50)
	2	2 (11.89)	1 (10.60)	1 (11.49)		3 (11.75)	2 (11.15)	
	3			2 (12.96)	4 (12.16)	1 (12.82)	4 (11.28)	4 (12.89)
	4	3 (13.96)		3 (13.79)		2 (13.44)		5 (13.68)
	5	4 (13.84)		5 (13.72)		5 (13.99)		
1ypi	**1**	**3 (0.65)**		**1 (0.73)**	**1 (0.99)**	**2 (0.88)**	**2 (0.68)**	**3 (0.60)**
	2		1 (7.58)	2 (10.99)		3 (11.46)		2 (9.38)
	3			5 (18.35)				4 (20.47), 5 (18.34)
	4	4 (17.45)	5 (15.87)		2 (16.64)			
	5					1 (5.36)	3 (6.22)	1 (5.23)
3ptn	**1**	**1 (0.39)**	**4 (0.55)**	**1 (0.54)**	**1 (0.30)**	**1 (0.26)**	**1 (0.45)**	
	2	3 (7.09)	5 (6.05)	5 (6.39)	2 (6.86)	5 (6.27)		
	3	2 (23.60)		2 (23.41)	3 (23.31)		3 (23.76)	
	4	5 (22.49)		3 (21.67)	5 (22.63)			
	5	4 (18.91)	2 (17.35)					2 (18.82)
	6			4 (12.06)		4 (12.33)	2 (12.95)	
2fbp	**1**	**1 (1.18)**	**3 (1.53)**	**1 (0.86)**	**1 (1.21)**, **2 (1.06)**		**5 (0.53)**	**2 (1.26)**
	2	**5 (1.44)**		**3 (0.72)**	**3 (1.23)**	**1 (0.44)**	**1 (1.09)**, **2(0.97)**	
	3	2 (11.94)		5 (12.83)		2 (13.66)		3 (11.14)
	4			2 (19.22)		4 (18.72)	3 (19.48)	1 (18.61)
	5				5 (24.85)	5 (24.89)	4 (23.61)	
	6	4 (19.53)			4 (18.93)			4 (19.65)
2had	**1**	**2 (0.63)**	**1 (0.55)**	**3 (0.80)**	**2 (0.93)**			
	2	1 (9.47)	2 (7.53)	1 (8.97)	1 (9.49)	1 (9.86)	1 (7.24)	2 (8.87)
	3	4 (10.79)	4 (10.44)	4 (10.35)	5 (10.78)			
	4	5 (29.35)		2 (29.25)	3 (29.33)	2 (29.36)	3 (29.65)	
	5	3 (9.78)	3 (8.95)			4 (10.52)		

Distance of each cluster center from the ligand in the X-ray structure of the corresponding bound enzyme after structural superposition.

[a] Consensus sites in the substrate binding region are shown in bold.

(Table 3, site 2), and no further discrimination between these two sites could be made based on the mapping results alone. For ribonuclease T_1 (9rnt), only phenol clusters at the binding site (not shown in Table 3), while there is a cluster with six different probes about 6 Å away (Table 3, site 1). In order to improve the reliability of the mapping results, we combined the results obtained by GRAMM and by CS-Map to determine if further resolution of the binding site can be achieved. For enolase (1ebh), the CS-Map based method adds six additional clusters of four different probes (Table 4, site 1) to the five probe clusters we have already obtained by GRAMM at the binding site. In contrast, for the second consensus site with five clusters (Table 3, site 2) only four probes were added on using CS-Map (Table 4, site 3). As a result, the known binding site of enolase was observed to have 11 clusters of five different solvents, the most clusters of any site on the protein. For ribonuclease T_1 (9rnt), the results were not as straightforward. While CS-Map added six new clusters at the binding site (Table 4, site 1) to the single phenol cluster that was given by the GRAMM-based mapping, and three more probes in the second sub-site (Table 4, site 5), these two sub-sites appear to be separate from each other, and hence we were not able to combine these locations into a single consensus site. As a result, the seven overlapping clusters in the binding site (Table 5, site 6) were still not enough to overcome a different site at which the combined algorithm yields ten clusters (Table 5, site 3). In addition, sites 1 and 2 in Table 5 show other locations with seven and nine clusters, respectively. Therefore, while CS-Map added a significant number of probes to the experimental binding site, they were not sufficient enough to compensate for the poor GRAMM result.

Comparison of bound and unbound structures of ribonuclease T_1 reveals that the side-chain of Y45 plays a dominant role in the binding of all ligands (see Figure 3b). As stated earlier, Y45 is considered as the "lid" of the binding site, and its side-chain takes on two different conformations. In the apo enzyme, Y45 forms a hydrogen bond

Table 4. Free energy ranking of probe clusters mapped by CS-Map

Protein	Con. site[a]	Probe					
		Acetone	Urea	Isopropanol	*t*-Butanol	Phenol	Acetonitrile
1ebh	**1**		**1** (**0.69**), **5** (**0.90**)		**1** (**0.39**)	**1** (**0.60**)	**1** (**0.92**), **3** (**1.34**)
	2	3 (21.31)		5 (21.39)	3 (20.79)	4 (20.46)	5 (20.65)
	3	1 (16.60)		1 (16.78)	5 (17.69)		2 (16.10)
	4	2 (13.75)	2 (11.96)	3 (14.09)			
	5	5 (9.46)	3 (9.460)	2 (9.63)			
	6				2 (18.43), 4 (22.28)	2 (22.47)	
9rnt	**1**	**2** (**0.56**)	**4** (**0.62**)	**3** (**0.29**)	**2** (**0.42**)	**5** (**0.26**)	**2** (**0.52**)
	2	1 (10.99)	3 (10.88)	2 (10.50)	1 (10.53)	2 (11.22), 4 (10.38)	1 (11.18)
	3	4 (12.06)	2 (11.85)	4 (11.68)	4 (12.67)		4 (12.73)
	4	3 (12.01)	1 (10.45)	1 (11.05)		1 (10.75)	
	5				**3** (**0.63**)	**3** (**0.41**)	**3** (**0.45**)

Distance of each cluster center from the ligand in the X-ray structure of the corresponding bound enzyme after structural superposition.

[a] Consensus sites in the substrate binding region are shown in bold.

with N99 and partially blocks the binding site from interacting with the substrate. The binding of a substrate analog or inhibitor moves Y45 to a different rotameric state, breaks the hydrogen bond with N99, and makes the pocket more accessible. We note that both CS-Map and GRAMM place some ligand clusters in the active site, even when mapping the apo structure, but other locations have higher numbers of different probes, resulting in false positives. However, the correct site can be easily selected if there is any information on its approximate position, and then the analysis of sub-clusters provides the required characterization of the binding site residues as for the other five proteins. Nevertheless, this result emphasizes the need for further testing of the mapping algorithm

Table 5. Free energy ranking of probe clusters mapped by either GRAMM and/or CS-Map for enolase (1ebh) and ribonuclease T_1 (9rnt) apostructures

Protein	Con. site[a]	Probe						
		Acetone	Urea	DMSO	Isopropanol	*t*-Butanol	Phenol	Acetonitrile
1ebh	**1**		**5** (**1.17**), **1**[b] (**0.69**), **5**[b] (**0.90**)	**3** (**0.72**)		**4** (**0.35**), **1**[b] (**0.39**)	**1** (**0.55**), **1**[b] (**0.60**)	**1** (**0.66**), **1**[b] (**0.92**), **3**[b] (**1.34**)
	2	1 (16.27), 1[b] (16.60)	4 (15.99)	2 (16.28)	1 (15.94), 1[b] (16.78)	5[b] (17.69)		2 (16.16), 2[b] (16.10)
	3	4 (23.13)		5 (23.02)	2 (22.18)			
	4			1 (7.82)		3 (7.38)	5[b] (8.42)	3 (8.13)
	5					1 (9.45)	2 (8.42), 3 (10.43)	
	6	2 (25.00), 4[b] (25.15)			4 (24.47), 4[b] (24.64)		4 (24.12)	
9rnt	1	5 (6.80)	4 (5.72), 5[b] (6.84)	4 (6.66)	5 (6.68)	4 (6.21)		3 (6.50)
	2	2 (11.89), 3[b] (12.01)	1 (10.60), 1[b] (10.45)	1 (11.49)	1[b] (11.05)	3 (11.75)	2 (11.15), 1[b] (10.75)	
	3	4[b] (12.06)	2[b] (11.85)	2 (12.96)	4 (12.16), 4[b] (11.68)	1 (12.82), 4[b] (12.67)	4 (11.28)	4 (12.89), 4[b] (12.73)
	4	3 (13.96), 5[b] (14.15)		3 (13.79)	5[b] (14.06)	2 (13.44)		5 (13.68)
	5	4 (13.84)		5 (13.72)		5 (13.99)		
	6[c]	**2**[b] (**0.56**)	**4**[b] (**0.62**)		**3**[b] (**0.29**)	**2**[b] (**0.42**)	**5**[b] (**0.26**) **5** (**0.29**)	**2**[b] (**0.52**)

Distance of each cluster center from the ligand in the X-ray structure of the corresponding bound enzyme after structural superposition.

[a] Consensus sites in the substrate binding region are shown in bold.

[b] Cluster centers derived from mapping with CS-Map. All cluster centers not denoted with [b] are derived from the GRAMM method.

[c] This consensus site corresponds to the experimental binding site for ribonuclease T_1.

in order to better understand the limits of its applicability. In particular, we expect that molecular dynamics simulation prior to docking would break the Y45–N99 hydrogen bond, making the binding site more accessible to the small probes.

Discussion

Why do organic solvents prefer binding at enzyme active sites?

The weakly specific binding of different ligands at the active sites of enzymes has been confirmed both by X-ray crystallography[4–8] and by NMR methods.[9] The results of computational solvent mapping show three properties of enzyme binding sites that might help to understand why such sites attract organic molecules, regardless of their sizes and polarities. First, the active sites of most enzymes are fairly large pockets that surround the small ligands, and thus provide a substantial number of ligand–protein contacts. Indeed, the intermolecular van der Waals energy is generally the largest contribution to the binding free energy. However, the van der Waals term may reach similarly low values in other clefts, far from the active site. Second, some fraction of this interface is non-polar and interacts with non-polar fragments of the ligand. The hydrophobic interactions provide another major contribution to the binding free energy. Again, it is important to note that similarly strong or even stronger non-polar contributions can occur in hydrophobic pockets that have nothing to do with the active site. Third, the net contributions of polar atoms to the free energy are relatively small, because the favorable electrostatic interactions are generally compensated by unfavorable desolvation of the partial charges. However, the presence of several polar patches in the active site is very important, because it enables the ligand to bind in a number of rotational/translational states. In each conformation, the polar parts of the ligand form one or two hydrogen bonds, or at least favorable electrostatic interactions, with one of the polar groups on the protein.[10] Due to the multiplicity of the bound conformations, a ligand binding in the active site retains more of its rotational/translational entropy than one that binds elsewhere in a single conformational state, and the resulting difference in the free energy makes tight binding in small crevices less favorable.

While good shape complementarity, substantial hydrophobic interactions, and the existence of several polar patches all seem to be necessary to steer small ligands toward the active site, the relative importance of these three factors remains uncertain, primarily because the entropic contributions are difficult to assess. Mapping generally shows three to eight bound states, corresponding to well-populated sub-clusters, in the active site. Assuming that these states are equally likely, accounting for the multiplicity would lower the free energy by 0.6–2.0 kcal/mol. Although this contribution is not very large, it may be important, because the van der Waals, electrostatic, and desolvation components on their own frequently give similar values for a number of crevices. Indeed, what primarily distinguishes the active site from other pockets in enzymes is not the size or hydrophobicity, but the existence of several polar groups that are always required for catalytic activity. Thus, the presence of catalytically important polar residues that can serve as acids and/or bases can additionally aid in providing a binding site with adequate polarity in discrete locations in order to fulfill the above requirement for substrate binding. Most of these groups can also form favorable (hydrogen bonding or electrostatic) interactions with the polar moiety of the ligand. Size also matters, as the site must be large enough to accommodate the small probes in multiple conformational states, each with good shape complementarity, resulting in favorable van der Waals interactions. If the active site is very large, as in the case of the hen-egg lysozyme,[9,10] the lowest free energy clusters of the different probes usually overlap in a sub-site of the active site (e.g. site C for lysozyme). As we have shown, additional consensus sites occur in other sub-sites for a number of enzymes, delineating further parts of the active site. We have found that mapping also works for non-enzyme proteins such as streptavidin that has a deep, partially hydrophobic binding site that includes a number of hydrogen bond donor and acceptor groups (unpublished results). It is not yet clear whether the method can be extended to recognize potential protein–peptide and protein–protein interactions sites that are substantially more planar than the sites considered here.

While the above analysis provides some insight on the origin of the weakly specific binding in the active site, our conclusions, including the importance of multiple bound conformations, are based on mapping results, and hence depend on the validity of the free energy evaluation models used in the calculations. However, the X-ray structures of the few proteins, determined in organic solvents, support the existence of multiple bound states.[5–8] It is tempting to speculate that attracting a large variety of ligands to the active site has evolutionary advantages, and retaining rotational/translational degrees of freedom is a good way to stabilize otherwise very weak complexes. In fact, enzymes with very broad substrate specificity such as cytochrome *P*450s that need to metabolize a wide range of xenobiotics, including chemicals produced by the modern chemical industry, bind some substrates in several conformations, resulting in a mixture of different metabolites.[48]

Most enzymes, however, bind their specific ligands (substrates, transition state analogs, and inhibitors) in unique, well-defined conformations, forming four to six hydrogen bonds. Although binding in a unique conformation implies that

more rotational/translational entropy is lost, the increase in the free energy is more than compensated by favorable van der Waals and electrostatic interactions.

Why do our mapping algorithms work?

Mapping proteins computationally rather than experimentally goes back as far as 1985, when Goodford developed the GRID program[49] to map receptor sites. Another popular approach to mapping is the Multiple Copy Simultaneous Search (MCSS†[50] version 2.1, Harvard University, Cambridge, MA, USA) method, which optimizes the free energy of numerous ligand copies simultaneously, each transparent to the others but subject to the full force of the receptor. However, the classical algorithms generally fail to reproduce the available NMR and X-ray data on the binding of organic solvents to proteins. The major problem is that they result in too many energy minima on the surface of the protein, and it is difficult to determine which of these minima is actually relevant.[4] This shortcoming was demonstrated by English *et al.*,[8] who used both GRID and MCSS to map thermolysin for the binding sites of isopropanol, acetone, acetonitrile, and phenol, and compared the results to those of mapping experiments. While they found local minima close to the experimentally observed binding positions, the closest minima were generally not among those with the lowest free energies, resulting in false positives (i.e. configurations with favorable energy which are not located near any experimentally observed binding site).

The algorithms presented here differ from traditional mapping methods in four major respects. First, while very different, both the CS-Map algorithm and the GRAMM-based mapping method provide much better sampling of the potential binding sites than GRID and MCSS that include only local minimization rather than any systematic search. Second, while neither GRID nor MCSS account for desolvation, the free energy potential used in step 2 of our mapping algorithm includes a relatively accurate electrostatics and desolvation model. Third, the docked ligand positions are clustered, and the clusters, rather than individual docked conformations, are ranked on the basis of their average free energies. The main goal of this step is to estimate the entropic effects of the multiple bound states, a contribution that otherwise would not be accounted for in our model. Discrimination by clustering, introduced by Baker and co-workers in the context of protein structure determination,[51] and extended by us to protein–protein docking,[52] is based on the idea that the native structure has more structural neighbors than other, non-native conformations do. Indeed, the multiple bound states in the active site result in a cluster of low energy conformations that define a relatively broad free energy minimum.

Clustering and considering the average free energies of the clusters eliminate most of the local minima that correspond to binding in narrow, isolated pockets. Nevertheless, this approach is unable to fully account for the extra entropy that comes from the multiplicity of the bound states, and hence the mapping is not expected to yield perfect discrimination of the correct bound states. In fact, the ligand positions that are the closest to the active site do not necessarily have the lowest values of the free energy (see Tables 1–4). The fourth, and probably the most important, difference between our method and earlier approaches is that we seek consensus sites at which the lowest free energy clusters of different solvents overlap. Restricting considerations to consensus sites implies that some false positives for specific ligands can be tolerated. For example, if the probability of obtaining a false positive is as high as 20%, but the false positives for the different probes are independently distributed over the protein surface, then the probability of obtaining a false consensus site using six probes is less than 0.01%. In reality, the situation is less favorable, since the false positives tend to be in relatively large pockets and hence are not independent. Nevertheless, as we have shown, mapping with six or seven probes usually gives very good results.

Solvent mapping generalizes the geometric analysis of protein binding sites

Since the size and shape of a protein cavity dictate the geometry of ligands that can bind there, geometry-based computational tools have been used to predict putative binding sites.[15–20] For example, PASS (Putative Active Sites with Spheres) is a simple computational tool that maps the protein surface with a water-sized sphere to characterize regions of buried volume. Computational solvent mapping generalizes the geometric analysis by using a number of small molecular probes with different sizes and shapes, and consisting of polar and non-polar fragments. As we argued, such probes are likely to prefer functional sites to other cavities, because the functional sites generally include a mixture of non-polar patches (for strong binding) and polar groups (for enzymatic activity). While it would be possible to use an arbitrary set of probes, the use of small organic solvents ensures that the results can be compared to data from solvent mapping experiments.[4–8] Furthermore, the detailed atomic models used in the mapping provide hydrogen bonding information that cannot be derived by a geometric analysis.

For comparison with the mapping results we applied the PASS algorithm[18] to the ligand-bound and apo structures of the six enzymes. As shown in Table 6, with two exceptions the ranked list of

† Evensen, E., Joseph-McCarthy, D. & Karplus, M. (1997).

Table 6. Prediction of experimental binding site locations using the PASS algorithm

Enzyme	Bound structure	Prediction of binding site	Ranking of successful prediction[a]	Apostructure	Prediction of binding site	Ranking of successful prediction[a]
Enolase	1ebg	Failed	–/12	1ebh	Succeeded	1/14
Trypsin	1tng	Succeeded	1/6	3ptn	Succeeded	3/5
Ribonuclease T_1	1rnt	Succeeded	1/3	9rnt	Succeeded	1/2
Triosephosphate isomerase	2ypi	Failed	–/9	1ypi	Succeeded	7/11
Fructose-1,6-bisphosphatase	1fbc	Succeeded	6/11	2fbp	Succeeded	8/14

[a] The number to the left represents the ranking of the binding site prediction while the number to the right represents the total number of predicted sites along the protein surface. Any failed attempt to predict the experimental binding site is represented by –.

putative binding sites, generated by PASS, includes the known binding site. However, the latter is ranked first (in degree of confidence) only for two of the bound structures (trypsin and ribonuclease T_1) and two of the apo structures (enolase and ribonuclease T_1). We recall that the mapping fails to rank the binding site first only for the ribonuclease T_1 apo structure. For haloalkane dehalogenase, PASS was able to find the location of one end (the entrance) of the extended binding channel, and was not able to successfully predict any additional part of the protein with known biological significance, including the active site (Table 7). Additionally, the prediction of the channel entrance for the apo and bound forms of the enzymes, respectively, were fifth and fourth ranked. It is clear that PASS, employing a single spherical probe, provides less specific information on the binding site than the mapping, which employs a variety of probes, each supplying its own geometric orientations as well as its interspersions of hydrophobic locations with hydrogen bond donor and acceptor atoms.

Mapping *versus* docking

In principle, one should be able to predict the binding properties of a protein by docking various ligands to its binding sites using docking programs such as DOCK[53] or Autodock.[54] However, ligand binding can substantially alter the structure of the protein (see, e.g. Yu & Koshland),[55] and docking to the unbound form is a non-trivial problem.[56,57] Low resolution and theoretically predicted receptors present even greater challenges.[58] Docking programs are frequently used for finding potential ligands in large databases, but it has been observed that the ligands found usually do not have the tightest fit, but instead leave some space movement in the binding site.[56] Small ligands are obviously easier to dock than large ones, and thus mapping is generally less sensitive to variations in structure than docking. In fact, the only protein for which mapping failed to correctly identify the residues that are important for ligand binding was ribonuclease T_1. Docking results are generally more sensitive to the conformational differences between bound and apo forms, and hence the bound forms are used for drug design whenever available.

Conclusions

Computational mapping employs small organic molecules to probe the surface of proteins. We have performed mapping calculations for seven enzymes. With the exception of haloalkane dehalogenase, which binds very small substrates in a narrow channel, the probes cluster in major subsites of the substrate-binding site. For haloalkane dehalogenase, the clusters occur at the two ends of the channel, but the smaller probes also find the

Table 7. Prediction of binding site locations on haloalkane dehalogenase using the PASS algorithm

Enzyme	Bound structure, 2dhc	Prediction of binding site	Ranking of successful prediction[a]	Apo-structure, 2had	Prediction of binding site	Ranking successful prediction[a]
Haloalkane dehalogenase	Active site	Failed	–/9	Active site	Failed	–/12
	Consensus site 2[b]	Failed	–/9	Consensus site 2[b]	Failed	–/12
	Consensus site 3[b]	Succeeded	5/9	Consensus site 3[b]	Succeeded	4/12
	Consensus site 4[b]	Failed	–/9	Consensus site 4[b]	Failed	–/12

[a] The number to the left represents the ranking of the binding site prediction while the number to the right represents the total number of predicted sites along the protein surface. Any failed attempt to predict the experimental binding site is represented by –.
[b] Refer to Table 2 for the designation of these specific consensus site locations. The significance of each of these four consensus sites is described in the text. Note that the consensus sites specified in this Table are based solely on mapping results using 2dhc.

active site. Residues that interact with many ligands (substrate analogs, products, and inhibitors) also tend to bind the probes with high frequency. Since the probe–residue interactions reflect the residue's role in binding rather than catalytic activity, the method can be used to characterize the substrate binding sites of enzymes.

As we have shown, results are slightly better when mapping ligand bound *versus* apo structures of enzymes, a situation well known in structure-based drug design. However, since it is easier to dock very small molecules to a rigid protein than larger ones, the sensitivity of results to moderate changes in the protein coordinates remains relatively low.

Here we were able to resolve a number of open problems. First, mapping results for seven enzymes strongly suggest that the binding of small organic compounds at the active site of enzymes is a general phenomenon. Since X-ray structures have been determined in organic solvents only for a few enzymes, extending the analysis to more proteins was absolutely necessary for any progress. Second, the results convincingly show that solvent mapping can be performed computationally rather than experimentally. Third, the analysis of mapping results enabled us to better understand why the small ligands bind in the active site, regardless of their sizes and polarities. Fourth, comparing the probe–protein interactions in the mapping results with the ligand–protein interactions, extracted from the X-ray structures of enzyme–ligand complexes shows that the mapping provides a detailed and reliable characterization of enzyme binding sites, and we hope that we will be able to apply it to a number of structures produced by structural genomics initiatives.

Methods

Computational mapping by the CS-Map algorithm

The five computational steps of the algorithm are as follows.[9,10]

Step 1: rigid body search

In the CS-Map algorithm[10,11] a multi-start simplex method is used to move the probes around the protein, starting from a number of evenly distributed points over the entire protein surface, i.e. no *a priori* assumption is made about the location of the binding site. The scoring function in the search is given by:

$$\Delta G_s = \Delta E_{elec} + \Delta G_{des} + V_{exc} \quad (1)$$

where ΔE_{elec} denotes the direct (Coulombic) part of the electrostatic energy, ΔG_{des} is the desolvation free energy, and V_{exc} is an excluded volume penalty term such that $V_{exc} = 0$ if the ligand does not overlap with the protein. The electrostatic energy is determined by the expression $\Delta E_{elec} = \sum_i \Phi_i q_i$, where q_i is the charge of the ith probe atom, and Φ_i is the electrostatic field of the solvated protein at that point.[59,60] The electric field Φ is calculated by a finite difference Poisson–Boltzmann (FDPB) method[59,60] using the CONGEN program.[61] Dielectric constants $\varepsilon = 4$ and $\varepsilon = 78$ are used for the protein and the solvent, respectively. We use the template partial charges provided by the Quanta program† (Molecular Simulations, Inc) for the probe molecules. The desolvation term, ΔG_{des}, is obtained by the Atomic Contact Potential (ACP) model,[62] an atomic level extension of the Miyazava–Jernigan potential.[63]

In this step, we also use an alternative approach based on the docking program GRAMM (global range molecular matching).[64,65] The program places the protein and the ligand molecule on separate grids, and performs an exhaustive six-dimensional search through the relative intermolecular translations and rotations using a very efficient Fast Fourier Transform (FFT) correlation technique and a simple scoring function that measures shape complementarity and penalizes overlaps. We have used 1.5 Å grid step for translations and 15° increments for rotations. A total of 1000 docked conformations were retained for refinement in step 2.

Step 2: minimization and re-scoring

Step 1 produces a large number of protein–ligand complexes at various local minima of ΔG_s. The free energy of each complex is minimized using the more accurate free energy potential:

$$\Delta G = \Delta E_{elec} + \Delta E_{vdw} + \Delta G^*_{des} \quad (2)$$

where ΔE_{vdw} denotes the receptor–ligand van der Waals energy, and the superscript in ΔG^*_{des} emphasizes that the desolvation term includes the change in the solute–solvent van der Waals interaction energy. The sum $\Delta E_{elec} + \Delta G^*_{des}$ is obtained by the Analytic Continuum Electrostatic (ACE) model,[66] as implemented in version 27 of CHARMM[67] using the parameter set from version 19 of the program. The minimization is performed using an adopted basis Newton–Raphson method.[67] During the minimization the protein atoms are held fixed, while the atoms of the probe molecules are free to move.

Step 3: clustering and ranking

The minimized probe conformations from step 2 are grouped into clusters based on Cartesian coordinate information. The method creates a number of clusters such that the maximum distance between a cluster's hub and any of its members (the cluster radius) is smaller than half of the average distance between all the existing hubs. We have slightly modified this algorithm by introducing an explicit upper bound $U = 4.0$ Å on the cluster radius. We retain only the clusters with more than T entries, where the threshold T is defined by the average clusters size, $T = m/n$ if $T < 20$, where m is the total number of probes and n is the number of clusters. Otherwise $T = 20$, i.e. clusters with more than 20 elements are always retained. For each retained cluster, we calculate the partition function $Q_i = \sum_j \exp(-\Delta G_j/RT)$, obtained by summing the Boltzmann factors over the conformations in the ith cluster only. The clusters are ranked on the basis of their average free energies $\langle \Delta G \rangle_i = \sum_j p_{ij} \Delta G_j$, where $p_{ij} = \exp(-\Delta G_j/RT)/Q_i$, and the sum is taken over the members of the ith cluster.

† QUANTA/CHARMM Program, Molecular Simulations Inc., Waltham, MA, USA.

Step 4: determination of consensus sites

Mapping is primarily used to find "consensus" sites at which many different probe molecules cluster. In order to find the consensus sites, we select the minimum free energy conformation in each of the five lowest average free energy clusters for each solvent. The structures are superimposed, and the position at which most probes of different types overlap is defined as the main consensus site. An additional clustering of probes close to the main consensus site is likely to indicate another sub-site of the active site, and hence the probes in the second cluster are added to those already in the consensus site.

Step 5: sub-cluster analysis

For each ligand, the cluster at the consensus site is further divided into sub-clusters based on probe orientations and free energies. The latter are included, because similar conformations with very different free energies usually have different mechanisms of binding (e.g. different hydrogen bonding interactions), and hence it is preferable to group them into different sub-clusters.[10,11] The sub-clusters of the ith cluster are ranked on the basis of the probabilities $p_{ij} = Q_{ij}/Q_i$, where Q_i is the sum of the Boltzmann factors over all conformations of the ith cluster, and Q_{ij} is obtained by summing the Boltzmann factors over the conformations in the jth subcluster only. Each subcluster with $p_{ij} > 0.05$ was represented by a single conformation. The LIGPLOT program[68,69] of Thornton and co-workers was used to find the non-bonded interactions and hydrogen bonds formed between each probe conformation and the protein. After counting all interactions, we have determined their distribution among the residues of the protein, as shown in Figures 1, 3, and 4.

Extracting intermolecular interactions from enzyme–ligand complexes

For each enzyme considered here we downloaded the structures of all complexes available in the RCSB PDB, including those of close homologues. Non-bonded interactions and hydrogen bonds have been determined using the SAS program†. For each amino acid residue we counted the number of interactions, and determined how these interactions distribute among the various residues, as shown in Figures 1, 3, and 4. Details of the interactions and a description of the roles of individual residues in each enzyme are available electronically‡.

Acknowledgements

We thank Professor Dagmar Ringe for useful discussions. This research has been supported by grants DBI-9904834 from the National Science Foundation, GM61867 and GM64700 from the National Institute of Health, and P42 ES07381 from the National Institute of Environmental Health Sciences.

† http://www.biochem.ucl.ac.uk/bsm/pdbsum/
‡ http://structure.bu.edu/

References

1. Bonanno, J. B., Edo, C., Eswar, N., Pieper, U., Romanowski, M. J., Ilyin, V. *et al.* (2001). Structural genomics of enzymes involved in sterol/isoprenoid biosynthesis. *Proc. Natl Acad. Sci. USA*, **98**, 12896–12901.
2. Erlandsen, H., Abola, E. E. & Stevens, R. C. (2000). Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites. *Curr. Opin. Struct. Biol.* **10**, 719–730.
3. Skolnick, J., Fetrow, J. S. & Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nature Biotechnol.* **18**, 283–287.
4. Mattos, C. & Ringe, D. (1996). Locating and characterizing binding sites on proteins. *Nature Biotechnol.* **14**, 595–599.
5. Ringe, D. & Mattos, C. (1999). Analysis of the binding surfaces of proteins. *Med. Res. Rev.* **19**, 321–331.
6. Allen, K. N., Bellamacina, C. R., Ding, X., Jeffery, C. J., Mattos, C., Petsko, G. A. & Ringe, D. (1996). An experimental approach to mapping the binding surfaces of crystalline proteins. *J. Phys. Chem.* **100**, 2605–2611.
7. English, A. C., Done, S. H., Caves, L. S., Groom, C. R. & Hubbard, R. E. (1999). Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins: Struct. Funct. Genet.* **37**, 628–640.
8. English, A. C., Groom, C. R. & Hubbard, R. E. (2001). Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Eng.* **14**, 47–59.
9. Liepinsh, E. & Otting, G. (1997). Organic solvents identify specific ligand binding sites on protein surfaces. *Nature Biotechnol.* **15**, 264–268.
10. Dennis, S., Kortvelyesi, T. & Vajda, S. (2002). Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl Acad. Sci. USA*, **99**, 4290–4295.
11. Kortvelyesi, T., Dennis, S., Silberstein, M., Brown, L., III & Vajda, S. (2003). Algorithms for computational solvent mapping of proteins. *Proteins: Struct. Funct. Genet.* **51**, 340–351.
12. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). The evolutionary trace method defines the binding surfaces common to a protein family. *J. Mol. Biol.* **257**, 342–358.
13. Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M. *et al.* (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261.
14. del Sol Mesa, A., Pazos, F. & Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302.
15. Hendlich, M., Rippmann, F. & Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in enzymes. *J. Mol. Graph. Model.* **15**, 359–363.
16. Liang, J., Edelsbrunner, J. & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884–1897.
17. Edelsbrunner, H., Facello, M. & Liang, J. (1998). On the definition and the construction of pockets in macromolecules. *Disc. Appl. Math.* **88**, 83–102.
18. Brady, G. P., Jr & Stouten, P. F. W. (2000). Fast predic-

tion and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* **14**, 383–401.
19. Ondrechen, M. J., Clifton, J. G. & Ringe, D. (2001). THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl Acad. Sci USA*, **98**, 12473–12478.
20. Gutteridge, A., Bartlett, G. J. & Thornton, J. M. (2003). Using a neural network and sparial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.* **330**, 719–734.
21. Laskowski, R. A., Luscombe, N. M., Swindells, M. H. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.
22. Matthews, B. W. (1988). Structural basis of the action of thermolysin and related zinc peptidases. *Accts Chem. Res.* **21**, 333–340.
23. Lipscomb, W. N. & Strater, N. (1996). Recent advances in zinc enzymology. *Chem. Rev.* **96**, 2375–2433.
24. Marie-Claire, C., Ruffet, E., Antonczak, S., Beaumont, A., O'Donohue, M., Roques, B. P. & Fournie-Zaluski, M. C. (1997). Evidence by site-directed mutagenesis that arginine 203 of thermolysin and arginine 717 of neprilysin (neutral endopeptidase) play equivalent critical roles in substrate hydrolysis and inhibitor binding. *Biochemistry*, **36**, 13938–13945.
25. Marie-Claire, C., Ruffet, E., Tiraboschi, G. & Fournie-Zaluski, M. C. (1998). Differences in transition state stabilization between thermolysin (EC 3.4.24.27) and neprilysin (EC 3.4.24.11). *FEBS Letters*, **438**, 215–219.
26. Reed, G. H., Poyner, R. R., Larsen, T. M., Wedekind & J. E., Rayment, I. (1996). Structural and mechanistic studies of enolase. *Curr. Opin. Struct. Biol.* **6**, 736–743.
27. Zhang, E., Brewer, J. M., Minor, W., Carreira, L. A. & Lebioda, L. (1997). Mechanism of enolase: the crystal structure of asymmetric dimer enolase—2-phospho-D-glycerate/enolase—phosphoenolpyruvate at 2.0 Å resolution. *Biochemistry*, **36**, 12526–12534.
28. Larsen, T. M., Wedekind, J. E., Rayment, I. & Reed, G. H. (1996). A carboxylate oxygen of the substrate bridges the magnesium ions at the active site of enolase: structure of the yeast enzyme complexed with the equilibrium mixture of 2-phosphoglycerate and phosphoenolpyruvate at 1.8 Å resolution. *Biochemistry*, **35**, 4349–4358.
29. Brewer, J. M., Glover, C. V. C., Holland, M. J. & Lebioda, L. (1998). Significance of the enzymatic properties of yeast S39A enolase to the catalytic mechanism. *Biochim. Biophys. Acta*, **1383**, 351–355.
30. Vinarov, D. A. & Nowak, T. (1999). Role of His159 in yeast enolase catalysis. *Biochemistry*, **18**, 12138–12149.
31. Brewer, J. M., Holland, M. J. & Lebioda, L. (2000). The H159A mutant of yeast enolase has significant activity. *Biochem. Biophys. Res. Commun.* **276**, 1199–1202.
32. Steyaert, J. (1997). A decade of protein engineering on ribonuclease T_1. Atomic dissection of the enzyme–substrate interactions. *Eur. J. Biochem.* **247**, 1–11.
33. Hubner, B., Haensler, M. & Hahn, U. (1999). Modification of ribonuclease T1 specificity by random mutagenesis of the substrate binding segment. *Biochemistry*, **38**, 1371–1376.
34. Balaji, P. V., Saenger, W. & Rao, V. S. R. (1993). Computer modeling studies on the binding of 2′,5′-linked dinucleotide phosphates to ribonucleotide T_1—influence of subsite interactions on the substrate specificity. *J. Biomol. Struct. Dynam.* **10**, 891–903.
35. Kumar, K. & Walz, F. G. (2001). Probing functional perfection in substructures of ribonuclease T-1: double combinatorial random mutagenesis involving Asn43, Asn44, and Glu46 in the guanine binding loop. *Biochemistry*, **40**, 7348–7357.
36. Alber, T. C., Davenport, R. C., Jr, Giammona, D. A., Lolis, E., Petsko, G. A. & Ringe, D. (1987). Crystallography and site directed mutagenesis of yeast triosephosphate isomerase: what can we learn about catalysis from a "simple" enzyme? *Cold Spring Harbor Symp. Quant. Biol.* **52**, 603–613.
37. Joseph, D., Petzko, G. A. & Karplus, M. (1990). Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop. *Science*, **249**, 1425–1428.
38. Norledge, B. V., Lamber, A. M., Abagyan, R. A., Rottmann, A., Fernandez, A. M., Filimonov, V. V. *et al.* (2001). Modeling, mutagenesis, and structural studies on the fully conserved phosphate-binding loop (loop 8) of triosephosphate isomerase: toward a new substrate specificity. *Proteins: Struct. Funct. Genet.* **42**, 383–389.
39. Perona, J. J., Hedstrom, L., Rutter, W. J. & Fletterick, R. J. (1995). Structural origins of substrate determination in trypsin and chymotrypsin. *Biochemistry*, **34**, 1489–1499.
40. Sprang, S., Standing, T., Fletterick, R. J., Stroud, R. M., Finer-Moore, J., Xuong, N. H. *et al.* (1987). The three-dimensional structure of Asn102 mutant of trypsin: role of Asp102 in serine protease catalysis. *Science*, **237**, 905–909.
41. Helland, R., Leiros, I., Berglund, G. I., Willassen, N. P. & Smalas, A. O. (1998). The crystal structure of anionic salmon trypsin in complex with bovine pancreatic trypsin inhibitor. *Eur. J. Biochem.* **256**, 317–324.
42. Choe, J., Fromm, H. J. & Honzatko, R. B. (2000). Crystal structures of fructose-1,6-bisphosphatase: mechanism of catalysis and allosteric inhibition revealed in product complexes. *Biochemistry*, **39**, 8565–8574.
43. Liang, J., Huang, S., Zhang, Y., Ke, H. & Lipscomb, W. N. (1992). Crystal structure of the neutral form of fructose-1,6-bisphosphatase complexed with regulatory inhibitor fructose-2,6-bisphosphate at 2.6 Å resolution. *Proc. Natl Acad. Sci. USA*, **89**, 2404–2408.
44. Ke, H., Zhang, Y. & Lipscomb, W. N. (1990). Crystal structure of fructose-1,6-bisphosphatase complexed with fructose 6-phosphate, AMP, and magnesium. *Proc. Natl Acad. Sci. USA*, **87**, 5243–5247.
45. Villeret, V., Huang, S., Fromm, H. J. & Lipscomb, W. N. (1995). Crystallographic evidence for the action of potassium, thallium, and lithium ions on fructose-1,6-bisphosphatase. *Proc. Natl Acad. Sci. USA*, **92**, 8916–8920.
46. Schanstra, J. P., Ridder, I. S., Heimeriks, G. J., Rink, R., Poelarends, G. J., Kalk, K. H. *et al.* (1996). Kinetic characterization and X-ray structure of a mutant of haloalkane dehalogenase with higher catalytic activity and modified substrate range. *Biochemistry*, **35**, 13186–13195.
47. Pikkemaat, M. G., Ridder, I. S., Rozeboom, H. J., Kalk, K. H., Dijkstra, B. W. & Janssen, D. B. (1999). Crystallographic and kinetic evidence of a collision complex formed during halide import in haloalkane dehalogenase. *Biochemistry*, **38**, 12052–12061.
48. Ogury, K., Yamada, H. & Joshimura, H. (1994). Regiochemitry of cytochrome P450 isozymes. *Annu. Rev. Pharmacol. Toxicol.* **34**, 251–279.
49. Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites

on biologically important macromolecules. *J. Med. Chem.* **28**, 849–875.
50. Miranker, A. & Karplus, M. (1991). Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins: Struct. Funct. Genet.* **11**, 29–34.
51. Shortle, D., Simons, K. T. & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.
52. Comeau, S. R., Gatchell, D., Vajda, S. & Camacho, C. J. (2003). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, in the press..
53. Gschwend, D. A., Good, A. C. & Kuntz, I. D. (1996). Molecular docking towards drug discovery. *J. Mol. Recognit.* **9**, 175–816.
54. Goodsell, D. S. & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct. Funct. Genet.* **8**, 195–202.
55. Yu, E. W. & Koshland, D. E., Jr (2001). Propagating conformational changes over long (and short) distances in proteins. *Proc. Natl. Acad. Sci. USA*, **98**, 9517–9520.
56. Kazlauskas, R. J. (2000). Molecular modeling and biocatalysis: explanations, predictions, limitations, and opportunities. *Curr. Opin. Chem. Biol.* **4**, 81–88.
57. DeVoss, J. J., Sibbesen, O., Zhang, Z. & Ortiz de Montellano, P. R. (1997). Substrate docking algorithms and predictions of the substrate specificity of cytochrome P450cam and its L244A mutant. *J. Am. Chem. Soc.* **119**, 5489–5498.
58. Wojciechowski, M. & Skolnick, J. (2002). Docking of small ligands to low-resolution and theoretically predicted receptor structures. *J. Comp. Chem.* **23**, 189–197.
59. Gilson, M. K. & Honig, B. (1988). Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins: Struct. Funct. Genet.* **4**, 7–18.
60. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144–1149.
61. Bruccoleri, R. E. (1993). Grid positioning independence and the reduction of self-energy in the solution of the Poisson–Boltzmann equation. *J. Comp. Chem.* **14**, 1417–1422.
62. Zhang, C., Vasmatzis, G., Cornette, J. L. & DeLisi, C. (1996). Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.* **267**, 707–726.
63. Miyazava, S. & Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
64. Vakser, I. A. (1995). Protein docking for low-resolution structures. *Protein Eng.* **8**, 371–377.
65. Vakser, I. A., Matar, O. G. & Lam, C. F. (1999). A systematic study of low-resolution recognition in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **96**, 8477–8482.
66. Schaefer, M. & Karplus, M. A. (1996). A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.* **100**, 1578–1599.
67. Brooks, B. R., Bruccoleri, R. E., Olafson, B., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* **4**, 197–214.
68. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1995). LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.* **8**, 127–134.
69. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.

Edited by J. Thornton

(Received 28 May 2003; received in revised form 10 August 2003; accepted 11 August 2003)