Exploring Structurally Conserved Solvent Sites in Protein Families

Christopher A. Bottoms,[1] **Tommi A. White,**[2] **and John J. Tanner**[1,2*]
[1]*Department of Chemistry, University of Missouri-Columbia, Columbia, Missouri*
[2]*Department of Biochemistry, University of Missouri-Columbia, Columbia, Missouri*

ABSTRACT **Protein-bound water molecules are important components of protein structure, and therefore, protein function and energetics. Although structural conservation of solvent has been studied in a few protein families, a lack of suitable computational tools has hindered more comprehensive analyses. Herein we present a semiautomated computational approach for identifying solvent sites that are conserved among proteins sharing a common three-dimensional structure. This method is tested on six protein families: (1) monodomain cytochrome *c*, (2) fatty-acid binding protein, (3) lactate/malate dehydrogenase, (4) parvalbumin, (5) phospholipase A_2, and (6) serine protease. For each family, the method successfully identified previously known conserved solvent sites. Moreover, the method discovered 22 novel conserved solvent sites, some of which have higher degrees of conservation than the previously known sites. All six families studied had solvent sites with more than 90% conservation and these sites were invariably located in regions of the protein with very high sequence conservation. These results suggest that highly conserved solvent sites, by virtue of their proximity to conserved residues, should be considered as one of the defining three-dimensional structural characteristics of protein families and folds. Proteins 2006;64:404–421.**
© 2006 Wiley-Liss, Inc.

Key words: bioinformatics; cytochrome *c*; fatty-acid binding protein; lactate dehydrogenase; malate dehydrogenase; parvalbumin; EF-hand; phospholipase A_2; Rossmann fold; serine protease; water molecules

INTRODUCTION

Water is the solvent of biological chemistry and so it is not surprising that water molecules underlie many fundamental biochemical processes, including protein folding, enzymatic catalysis, and biomolecular recognition. The major driving force of protein folding, the hydrophobic effect, is an entropic gain associated with the solvent. Catalytic water molecules poised in enzyme active sites directly participate in the bond-breaking and bond-making steps of many enzymatic reactions. Other water molecules in active sites may be involved in binding and stabilizing the substrate during catalysis. Water is critical for biomolecular recognition and association because molecular surfaces must be desolvated during complex formation, and this shedding of solvent contributes significantly to the free energy of association. Lastly, and most relevant to our work, ordered water molecules that are tightly bound to proteins mediate intramolecular interactions within proteins and intermolecular interactions within biomolecular complexes.

The role of water as a mediator of noncovalent interactions is supported by high-resolution X-ray crystallographic data. Analysis of protein crystal structures has shown that water-mediated hydrogen bonds are abundant in protein complexes with DNA, protein, and small molecule ligands. It is estimated that 40% of all protein-DNA hydrogen bonds are water mediated,[1] and that protein–protein interfaces contain an average of 22 water molecules and 11 water-mediated hydrogen bonds.[2] There are many examples of protein–ligand complexes in which water molecules bridge the protein and a bound small molecule ligand. For example, our study of Rossmann dinucleotide-binding domains revealed that 30% of the hydrogen bonds between the protein and adenine dinucleotide cofactors (NAD(P), FAD) were water mediated.[3] Similarly, Babor et al.[4] showed that water-mediated hydrogen bonds were important for recognition of the ribose moieties of adenosin triphosphate (ATP), adenosine diphosphate (ADP), and flavin adenine dinucleotide (FAD).

Importantly, several examples from the structure-based drug design and medicinal chemistry literature have demonstrated the essential nature of water-mediated hydrogen bonds in protein-inhibitor recognition, including the class C β-lactamase AmpC,[5] HIV-1 integrase,[6] HIV pro-

Abbreviations: HIV, human immunodeficiency virus; PDB, Protein Data Bank; NMR, nuclear magnetic resonance; SSM, secondary structure matching; RT, room temperature; PQS, Protein Quaternary Structures; MSD, Macromolecular Structure Database; SD, standard deviation.

Grant sponsor: Interdisciplinary Plant Group and the Genetics Area Program of the University of Missouri-Columbia; Grant sponsor: National Library of Medicine; Grant number: 2-T15-LM07089-14.

Christopher A. Bottoms' present address is Department of Health Management and Informatics, University of Missouri-Columbia, Columbia, MO 65211.

*Correspondence to: John J. Tanner, University of Missouri-Columbia, Department of Chemistry, Columbia, Missouri 65211. E-mail: tannerjj@missouri.edu

Received 2 May 2005; Revised 16 December 2005; Accepted 24 February 2006

Published online 12 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21014

tease,[7,8] cyclin-dependent kinase-2,[9] Factor Xa,[10] thrombin,[11–13] herpes simplex virus type I thymidine kinase,[14] thymidylate synthase,[15,16] neuraminidase,[17,18] heat-labile enterotoxin,[19] and FKBP12.[20] In these cases, water molecules have a crucial role in ligand recognition by mediating hydrogen bonds between the ligand and the receptor protein. The overarching conclusion from this body of work was that successful drug design often requires explicit consideration of ordered water molecules in the binding pocket.

Given that protein-bound water molecules are important for maintaining proteins in their native conformations and for mediating biomolecular associations, there is considerable interest in elucidating exactly which water molecules might be particularly critical for protein structure and function. One hypothesis is that structurally and functionally important water molecules occupy solvent sites that are highly conserved among proteins sharing a common three-dimensional fold or active site structure. Conserved solvent sites have been examined only in a few protein families, such as fatty-acid binding proteins,[21] cytochrome *c*,[22] lectins,[23] phospholipase A_2,[24] ribonucleases,[25] Rossmann dinucleotide-binding proteins,[3] serine proteases,[13,26,27] and parvalbumins.[28]

As part of our ongoing interest in probing the roles of solvent in protein structure and function, we have developed a structural bioinformatics methodology for identifying and analyzing conserved solvent sites in protein structures. The method is tested on six protein families for which conserved solvent sites have been previously identified: monodomain cytochrome *c*, fatty-acid binding protein, lactate/malate dehydrogenase, parvalbumin, phospholipase A_2, and serine proteases. Our method not only identified known conserved solvent sites, but also revealed 22 novel ones. Analysis of the novel sites generated new hypotheses about protein structure, stability, and function. Furthermore, all six families had solvent sites with more than 90% conservation and these sites were invariably located in regions of the protein with very high sequence conservation. This result suggests that highly conserved water molecules should be considered as defining features of protein families, in addition to the better-known primary, secondary, and tertiary structural features of the polypeptide chain. Because our method has the potential to become fully automated, it could eventually be used to perform comprehensive analyses of conserved solvent sites in all protein families represented in the PDB.[29]

RESULTS

Description of Methodology

Our method is based on identifying equivalent water molecules in a superimposed set of structurally related proteins. For example, in the cytochrome *c* data set, the superimposed structures are shown in Figure 1(a). It is readily apparent that identification of conserved water molecules by manual inspection is an extremely challenging task and that robust computational methods are preferred, particularly when considering protein families with many members. Note that our method is structure-based, allowing for the comparison of proteins that have similar structures despite limited sequence similarities.

Equivalence of water molecules is determined by considering two criteria: (1) the spatial locations of water molecules relative to the protein, and (2) the noncovalent interactions that water molecules form with the protein. Spatial equivalence is assessed by calculating the distribution of water molecules within a nonredundant set of superimposed protein structures. This distribution is represented as a pseudo-electron density map, which is calculated from the water molecules of the superimposed family of structures. For example, the water distribution map for cytochrome *c* is shown in Figure 1(b). Note that there are several prominent features in the map. These regions of high electron density represent potential conserved solvent sites, and the map provides an intuitive visual representation of the structural context of these sites. One can see, for example, that there are several strong features near the heme group in this case [Fig. 1(b)]. Water molecules within 2.0 Å of these peaks are considered spatially equivalent.

Interaction equivalence is assessed by comparing the noncovalent interactions that each water molecule in the family makes with its respective protein. This criterion is quite general, and could include, for example, hydrogen bonding based on angle and distance cutoffs, and van der Waals interactions. For this initial study, a simple hydrogen-bonding criterion based on a 3.2 Å distance cutoff was used.

Water molecules in different structures of the family are considered structurally equivalent if (1) they are close to the same peak in the water distribution map, and (2) they have at least one interaction with the protein in common. This analysis leads to a quantitative measure of conservation defined as the percentage of structures in the protein family possessing an equivalent water molecule in a given solvent site. For example, the strongest peak in the cytochrome *c* map had a density value of 33σ and the corresponding site had 93% conservation (Table I, site 1). Thus, 93% (13/14) of the cytochrome *c* structures surveyed had an equivalent water molecule in this solvent site. Note that the percent conservation used here is analogous to the percent conservation of an individual amino acid residue within a multiple sequence alignment.

Cytochrome *c*

Using a yeast mitochondrial cytochrome *c* (1YCC)[30] as a query structure, we obtained a nonredundant cytochrome *c* data set consisting of 14 structures with resolution of at least 2.2 Å (see Materials and Methods). These matching structures were all monodomain cytochrome *c* proteins, comprising five eukaryotic mitochondrial cytochrome *c* proteins, seven bacterial cytochrome c_2 proteins, one bacterial cytochrome c_H, and one bacterial cytochrome c_{552}.

The top three conserved solvent sites had 86–93% conservation and peak heights in the water distribution map of 30–33σ (Table I, sites 1–3), whereas site 4 had lower conservation (64%). All four listed sites are near the

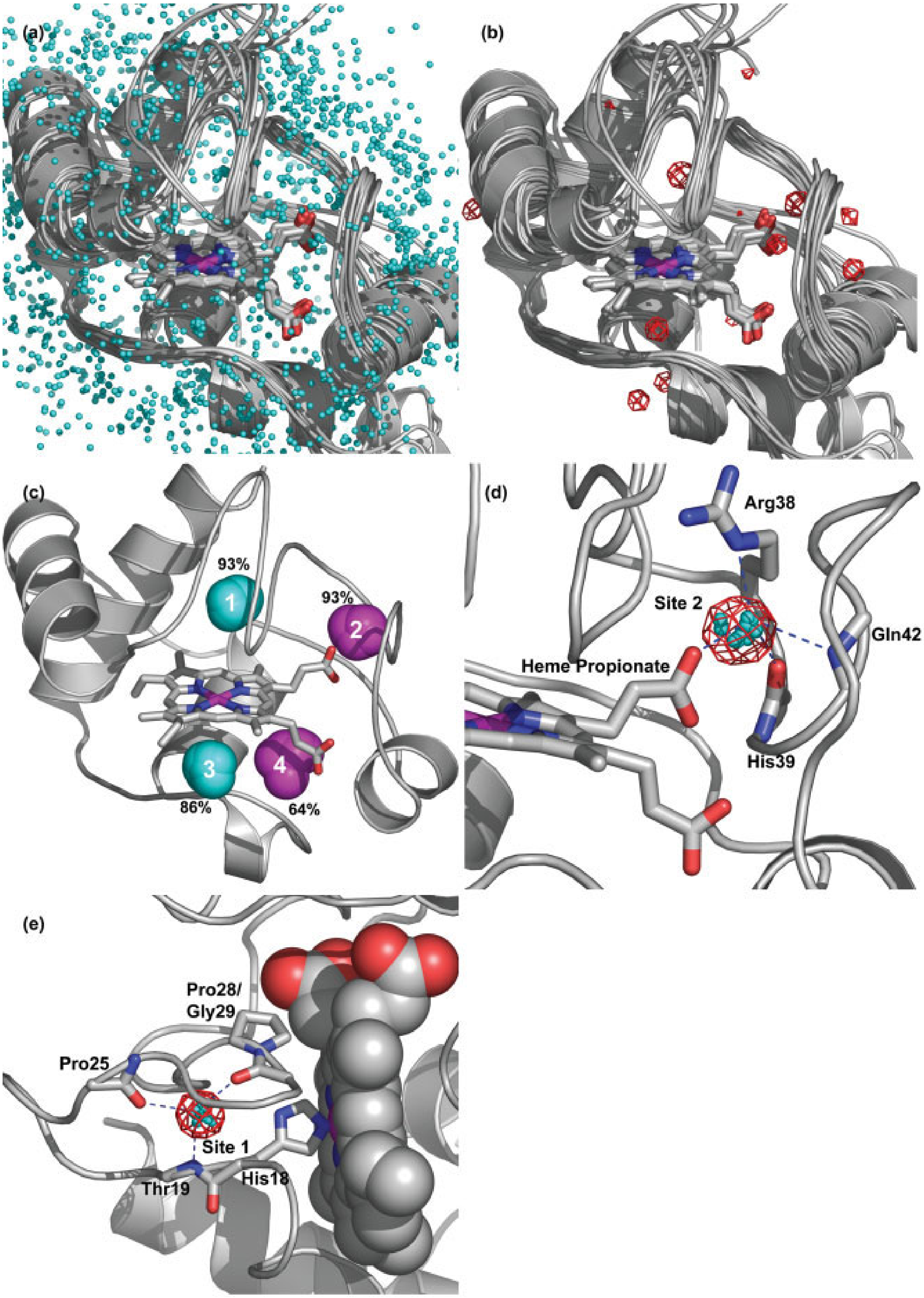


Fig. 1. Structurally conserved water molecules in cytochrome *c*. **a:** Superimposed cytochrome *c* structures including water molecules drawn as small cyan spheres. **b:** Superimposed structures and water-density map contoured at 15σ. The heme is drawn as sticks with Fe shown in magenta. **c:** Yeast iso-1-cytochrome *c* (1YCC) with water molecules at conserved sites 1–4. Cyan spheres denote novel solvent sites discovered by our methodology (sites 1, 3) and magenta spheres denote sites identified by our method that have also been previously reported in the literature (sites 2, 4). The percentage conservation values of the sites are indicated. **d:** Close-up view of site 2, which bridges the heme propionate with an Arg side-chain and backbone atoms. Water molecules belonging to this site are drawn as small cyan spheres and the water-density map is contoured at 10σ. **e:** Conserved solvent site 1 in the first Ω-loop. Water molecules belonging to this site are drawn as small cyan spheres and the water-density map is contoured at 10σ. The heme is drawn as spheres. Figures 1–6 were created by using PyMOL.[89]

TABLE I. Selected Conserved Solvent Sites of Cytochrome *c*

Site	Novel site?	Conservation (%)	Density peak (σ)	Representative interactions[a]		Water ID[a]
				Backbone	Side-chain	
1	√	93	33	N-Thr19 O-Pro25 O-Gly29		110
2		93	32	O-His39 N-Gln42	NE-Arg38 O1A-Hem104	121
3	√	86	30	O-Lys79 N-Ala81		122
4		64	20		ND2-Asn52 OH-Tyr67 OG1-Thr78	166

[a] Representative interactions and water ID for structure with PDB code 1YCC.

heme [Fig. 1(c)], which suggests possible roles for these water molecules in stabilizing protein structural elements that bind the heme.

Conservation of sites 2 and 4 in this family has been discussed previously.[22,31,32] Thus, our method successfully identified known conserved solvent sites. Site 2 has been described by Benning et al.[33] as conserved in eukaryotic mitochondrial cytochrome *c* and in bacterial cytochrome c_2. This site is interesting because it bridges the heme propionate with the surrounding protein [Fig. 1(d)]. This interaction likely helps to stabilize the heme and orient it correctly in the active site.

In addition, our method identified two novel sites with 93 and 86% conservation (Table I, sites 1 and 3). A water molecule located in site 1 has been discussed separately for cytochrome c_2 from *Rhodopila globiformis*,[33] *Rhodopseudomonas palustris*,[34] and *Paracoccus denitrificans*.[35] To our knowledge, the wider conservation of this site among monodomain cytochrome *c* structures has not been previously appreciated. This site is located in the middle of the first Ω-loop [Fig. 1(e)], and the water molecule is buried by the protein backbone. Within members of our data set, site 1 typically forms a bridge between a backbone amino group and two backbone carbonyls within the middle of this loop [Fig. 1(e)].

It is interesting to note that this conserved site exists despite variations of sequence position or spatial positions, of the solvent-coordinating residues. For example, the length of this Ω-loop is variable in our data set [Fig. 1(b)]. More specifically, given that the amino interaction is with residue number *i* of the sequence, the first carbonyl interaction is located at a position within $i + 3$ to $i + 6$ and the second carbonyl interaction is located at a position within $i + 7$ to $i + 13$.

Despite the structural variation of the Ω-loop, the loop does contain a conserved Gly-Pro motif that interacts with the conserved water molecule. The second interacting carbonyl (i.e., at a position within $i + 7$ and $i + 13$) belongs to the Gly residue of the Gly-Pro motif. Therefore, this carbonyl lies in the same plane as the N, C_α, and C_β atoms of the proline. The conserved Gly-Pro structural feature of the Ω-loop probably contributes to the observed high level of conservation of solvent site 1.

We note that site 1 is located near residues that have 100% sequence identity in our cytochrome *c* data set. Site 1 interacts with Gly of the conserved Gly-Pro motif and it is very close (4.2 Å in 1YCC, for example) to the His residue that binds the heme [Fig. 1(e)]. These three residues are among only 12 residues with 100% sequence identity in our data set. Site 2 is likewise situated near a conserved feature of the protein. Site 2 forms a hydrogen bond to a propionate of the heme cofactor [Fig. 1(d)]. Obviously, the heme is a conserved and essential feature of cytochrome *c*.

Fatty-Acid Binding Proteins

Prendergast's group[21,36] has extensively studied a conserved, internal solvent site in fatty-acid binding proteins using NMR, molecular dynamics simulations, and analysis of crystal structure data. In our study, this site corresponded to the top peak in the water distribution map (62σ), and the site had 100% conservation [Table II, Fig. 2(a)]. Thus, as in the monodomain cytochrome *c* case, our method successfully identified a known conserved solvent site.

In addition, our method identified two novel sites with more than 80% conservation. Site 3 is near the Prendergast site [Fig. 2(a)], and site 2 is near the bound fatty acid. Site 2 is particularly interesting because it seems to have a role in fatty-acid binding [Fig. 2(b)]. It typically donates hydrogen bonds to backbone carbonyl oxygen atoms that are *i* and $i + 3$ with respect to each other. These two residues lie at the end of the second helix of a pair of helices that have been described as a "lid" for the fatty-acid binding site.[37] In three structures, this structurally conserved water accepts a hydrogen bond from a conserved arginine residue (Arg126 in 1HMT) that, in turn, interacts with a negatively charged group of the bound fatty acid [Fig. 2(b)].

The two sites with more than 90% conservation are located near residues that have 100% sequence conservation. Site 1 is located between conserved residues Gly67 and Phe70 (1HMT numbering). In 1HMT, for example, the site 1 water molecule is 3.7 and 3.6 Å, respectively, from Gly67 and Phe70. Site 2 water molecules form van der Waals contacts with a conserved Phe [3.8 Å in 1HMT, see Fig. 2(b)]. We note that these three residues are among only four with 100% sequence identity in our data set.

TABLE II. Selected Conserved Solvent Sites of Fatty-Acid Binding Proteins

Site	Novel site?	Conservation (%)	Density peak (σ)	Representative interactions[a] Backbone	 Side-chain	Water ID[a]
1		100	62	O-Lys65		143
				O-Val68		
				N-Val84		
2	√	91	29	O-Ala33		135
				O-Thr36		
3	√	82	35	N-Lys65		
				O-Val68		212

[a] Representative interactions and water ID for structure with PDB code 1HMT.

Lactate/Malate Dehydrogenase

The lactate/malate dehydrogenase family features a structurally conserved water molecule in the dinucleotide-binding site. This site was previously described in the lactate dehydrogenase family[38,39] and, more generally, in proteins containing the dinucleotide-binding Rossmann fold.[3] This water molecule binds to the conserved Gly-rich loop of the Rossmann fold. In our study, this site corresponded to the top peak in the density map and was 100% conserved [Fig. 3(a); Table III, site 1).

Site 3, one of the interesting novel sites, had 79% conservation. Water at this site often serves as a salt bridge link, forming hydrogen bonds to both Arg and Asp residues [Fig. 3(b)]. These residues are located at positions i and $i + 3$ with respect to each other and are 100% conserved by sequence. It is well known that, within this family, the positively charged guanidinium group of this arginine stabilizes the negatively charged carboxylate of the substrate.[35] Thus, we see a correlation between amino acid sequence conservation and solvent structural conservation. The water-mediated salt-bridge occurs in 8 of the 14 proteins studied. In three of the remaining proteins, a water molecule near the peak interacts with either the arginine or aspartate, but not both. Using the Electron Density Server,[40] we observed that in another structure (1HYH) there was experimental density in both subunits that could be attributed to this conserved water. There was a sulfate ion in 6LDH and a crystal contact in 1LLD that could be attributed to disrupting the normal interactions.

An analysis on the fold level for NAD(P)-binding Rossmann-fold proteins (n = 126 structures) was also performed (see Materials and Methods) to test our method with a set of structures with low overall sequence identity. The 126-structure data set had mean pairwise sequence identity of only 14%. Nonetheless, our method clearly showed that the most conserved site on the fold level corresponded to site 1 of the LDH/MDH family (Table III, site 1). This result agrees with our previous survey of water molecules in Rossmann dinucleotide-binding domains, which was performed using laborious manual inspection of superimposed structures.[3] This highly conserved solvent site is critical for cofactor recognition because it bridges the dinucleotide pyrophosphate with the Gly-rich loop of the Rossmann fold.[3]

Parvalbumins

Parvalbumins are soluble EF-hand calcium-binding proteins containing two Ca^{2+}-binding sites: the CD-loop and the EF-loop. Conservation of solvent sites in the parvalbumin family has been extensively studied,[28,41–43] and thus represents a good test case for our method. Our calculation revealed many sites with conservation values >70% that corresponded to previously identified sites [Fig. 4; Table IV, sites 1–3, 7, 9, 12, 16–19). Nine novel solvent sites were also found (Table IV). Sites 4–6 are particularly notable because they have conservation values of 100%, yet they have not been mentioned, to our knowledge, as being conserved in this family.

Several conserved solvent sites are close to the Ca^{2+}-binding sites, and some of these water molecules have direct or indirect roles in binding Ca^{2+}. In this family of proteins, the ligands to the calcium ion are labeled with a letter corresponding to each of the three Cartesian coordinate axes and a positive or negative sign to indicate whether the ligand is "above" or "below" a specific plane. Within this coordination sphere, therefore, the −x ligand lies on the x axis and below the yz plane. Hence, there are +x, −x, +y, −y, +z, and −z ligands.[44] In the present study, site 3 is the −x ligand for the EF Ca^{2+}-binding site. Two other known conserved water molecules, sites 18 and 19,[43] interact with the +x ligand of the EF- and CD-loops, respectively. Site 8, a novel site with 86% conservation, forms a hydrogen bond with the CD −z ligand.

Novel site 6 (100% conservation) is particularly interesting because it has an indirect role in binding Ca^{2+} at the CD-loop [Fig. 4(b)]. This water molecule forms a hydrogen bond to the side-chain of residue 59, which is Asp in 1RRO or Glu in the remaining structures. One of the carboxylate oxygen atoms of Glu59 is the −x ligand of the Ca^{2+}-binding site and the other forms a hydrogen bond to the hydroxyl of Ser55, which is the +z ligand.[43] In 1RRO, however, the carboxylate oxygen atoms of Asp59 each use a water molecule extension to allow them to fill the same roles as in Glu59. One water serves as the −x ligand to the Ca^{2+} ion,[43] and the site 6 water forms a hydrogen bond with Ser55. These two water molecules are also within hydrogen-bonding distance of each other. Thus, we see that despite some variation in the Ca^{2+}-binding site, the site-6 water molecule still interacts with the −x ligand.

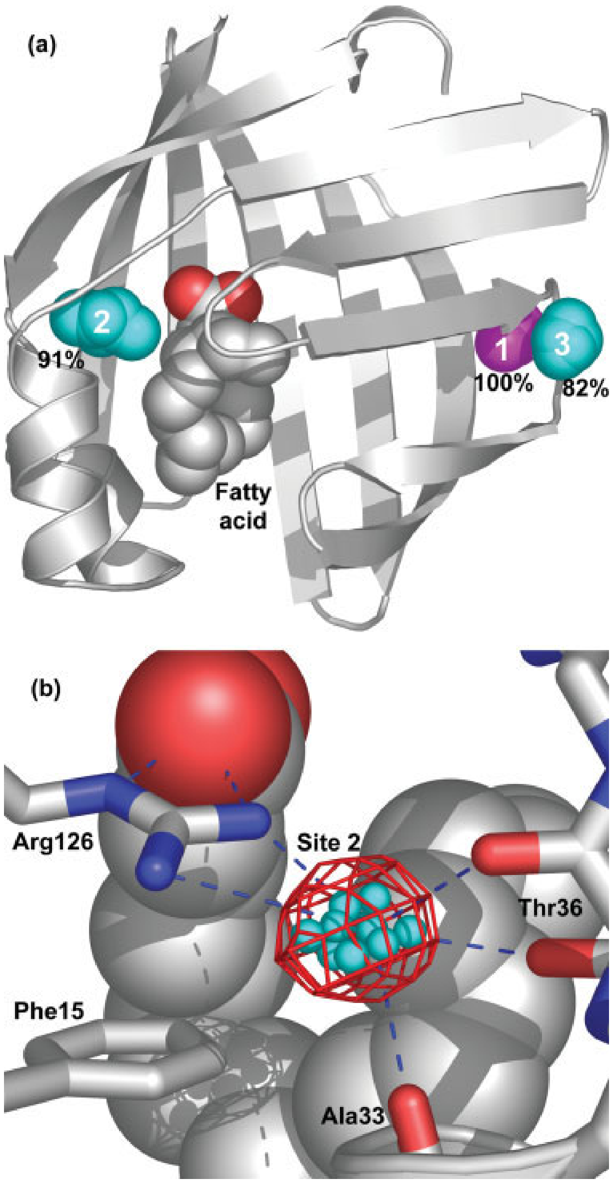


Fig. 2. Structurally conserved water molecules in fatty-acid binding proteins. **a:** Human muscle fatty-acid binding protein (1HMT) with water molecules from the superposition at conserved sites 1–3 (Table II). Cyan spheres denote novel solvent sites discovered by our methodology (sites 2, 3) and magenta spheres denote a site identified by our method that has also been previously reported in the literature (site 1). The percentage conservation values of the sites are indicated. **b:** Close-up view of conserved solvent site 2 (Table II). Water molecules belonging to this site are drawn as small cyan spheres and the water-density map is contoured at 10σ. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

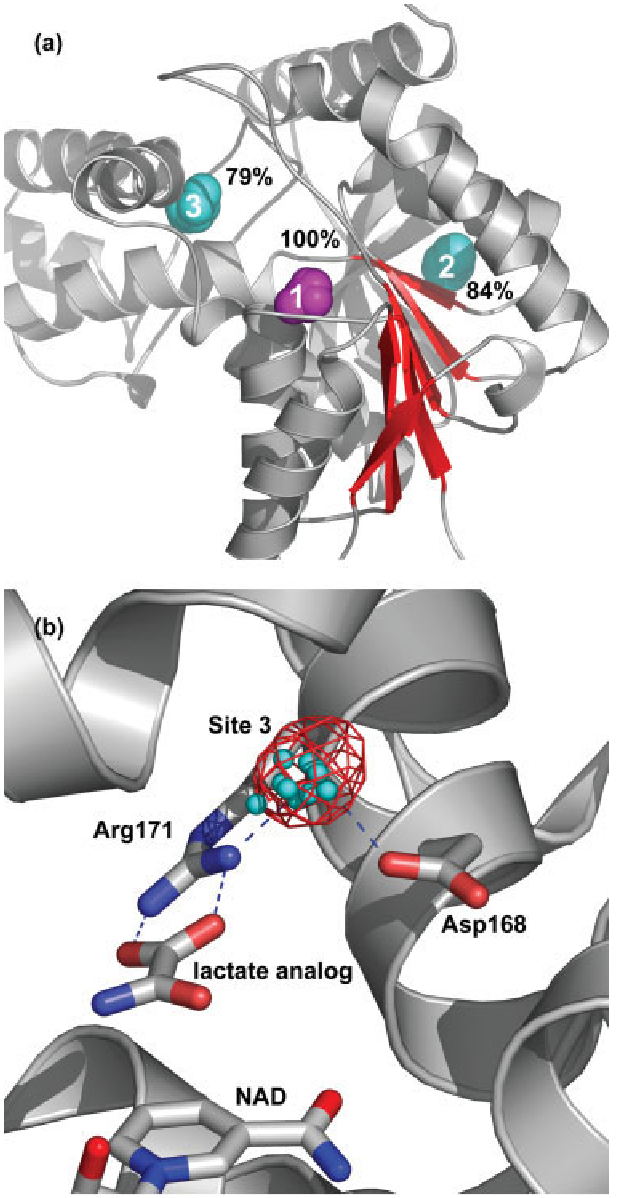


Fig. 3. Structurally conserved water molecules in lactate/malate dehydrogenases. **a:** Lactate dehydrogenase of *Plasmodium falciparum* (1LDG) with highly conserved water molecules from the superposition. Cyan spheres denote novel solvent sites discovered by our methodology (sites 2, 3) and magenta spheres denote a site identified by our method that has also been previously reported in the literature (site 1). The percentage conservation values of the sites are indicated. **b:** Close-up view of conserved solvent site 3, which is in the active site (Table III, site 3). Water molecules belonging to this site are drawn as small cyan spheres and the water-density map is contoured at 10σ. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

We also performed an analysis of a nonredundant set of proteins ($n = 38$) representing the EF-hand superfamily, as defined by SCOP.[45] The three sites with the highest conservation values (71–76%) were equivalent to sites 3, 16, and 17 of the parvalbumin family. In a family mostly composed of Ca^{2+} binding proteins, it is not surprising that site 3, the $-$x ligand of the EF-loop calcium ion, should be conserved. However, sites 16 and 17 are not obvious candidates for very high conservation. As noted by Strynadka and James[43] more than 15 years ago, these solvent sites effectively extend the β-sheet between the calcium-binding loops. They also mentioned that loops lacking calcium tended to form direct hydrogen bonds instead of water-mediated hydrogen bonds at these positions within the sheet. Their observations were made in a study that included five proteins: carp parvalbumin, tur-

TABLE III. Selected Conserved Solvent Sites of Lactate/Malate Dehydrogenases

				Representative interactions[a]		
Site	Novel site?	Conservation (%)	Density peak (σ)	Backbone	Side-chain	Water ID[a]
1		100	62	N-Gly29 N-Gly32	OG1-Thr97 NO2-NAD401[b]	13
2	√	86	39	O-Lys157 O-Ile160 N-Leu274	OG1-Thr273	11
3	√	79	35	O-Thr232	OD1-Asp168 NH1-Arg171	59

[a] Representative interactions and water ID for structure with PDB code 1LDG.

[b] NO2 is a pyrophosphate oxygen of the nicotinamide half of NAD. It is equivalent to atom name O2P in Schultze and Feigon.[97]

key troponin C, chicken troponin C, bovine calmodulin, and bovine intestinal calcium-binding protein. In our study, all three sites are also seen in frequenin (1G8I),[46] osteonectin (1SRA),[47] the calcium-binding pollen allergen Phl p 7 (1K9U),[48] S100 proteins (1E8A,[49] 1MHO,[50] 1IRJ[51]), and sarcoplasmic calcium-binding protein (2SCP).[52] One or more of the three top sites are also seen in many of the other proteins within the EF-hand superfamily.

Conservation of the β-sheet-extending solvent sites (sites 16 and 17 of parvalbumin) is particularly interesting in Phl p 7 (1K9U). In contrast to the other calcium-binding proteins, the two calcium-binding sites that form the β-sheet are from two different polypeptide chains [Fig. 4(c)]. Thus, these highly conserved water molecules extend an *inter*molecular β-sheet in Phl p 7, in contrast to an *intra*molecular β-sheet in the other proteins. Figure 4(c) depicts Phl p 7 with the three most highly conserved water molecules of the EF-hand superfamily. Despite this unique structural difference, this protein contains all three of the most conserved solvent sites of this family. This is an excellent example of structural conservation that would not likely be predicted by amino acid sequence alone.

Phospholipase A_2

In the phospholipase A_2 family, two known conserved solvent sites are part of an important hydrogen-bonding network that typically involves the N-terminal residue (e.g., Ser1), His48, Tyr52, Ser68, and Asp99.[24,53,54] One is considered to have more of a structural role (Table V, site 3), whereas the other has an important catalytic role (Table V, site 9). The structural water molecule typically forms hydrogen bonds to all of these residues except for His48. His48 interacts with Asp99 and with the catalytic water molecule. The structural water is 100% conserved in our data set and the catalytic water is 67% conserved [Fig. 5 and Table V, sites 3 and 9). In two cases (1HN4[55] and 1KVO[56]), the catalytic water was displaced by an inhibitor.

Interestingly, our method uncovered two 100% conserved novel solvent sites (Table V, sites 1 and 2) nestled between two α-helices [Fig. 5(a,b)]. The two helices are linked by two 100% conserved disulfide bonds (Cys1044—Cys1105, Cys1051—1098), which are always separated by two helical turns [Fig. 5(b)]. The pair of conserved water molecules bridges backbone carbonyls of the two disulfide links. These water molecules likely fulfill a structural role, and their ubiquity among members of this family invites further study. Sites 5 and 6 are also near conserved disulfide bonds. Site 5 (92% conserved) is within 4.0 Å of the previously mentioned Cys1105 [Fig. 5(b)] and site 6 (92% conserved) forms a hydrogen bond with the amino group of Cys1084, which forms a disulfide linkage with Cys1096.

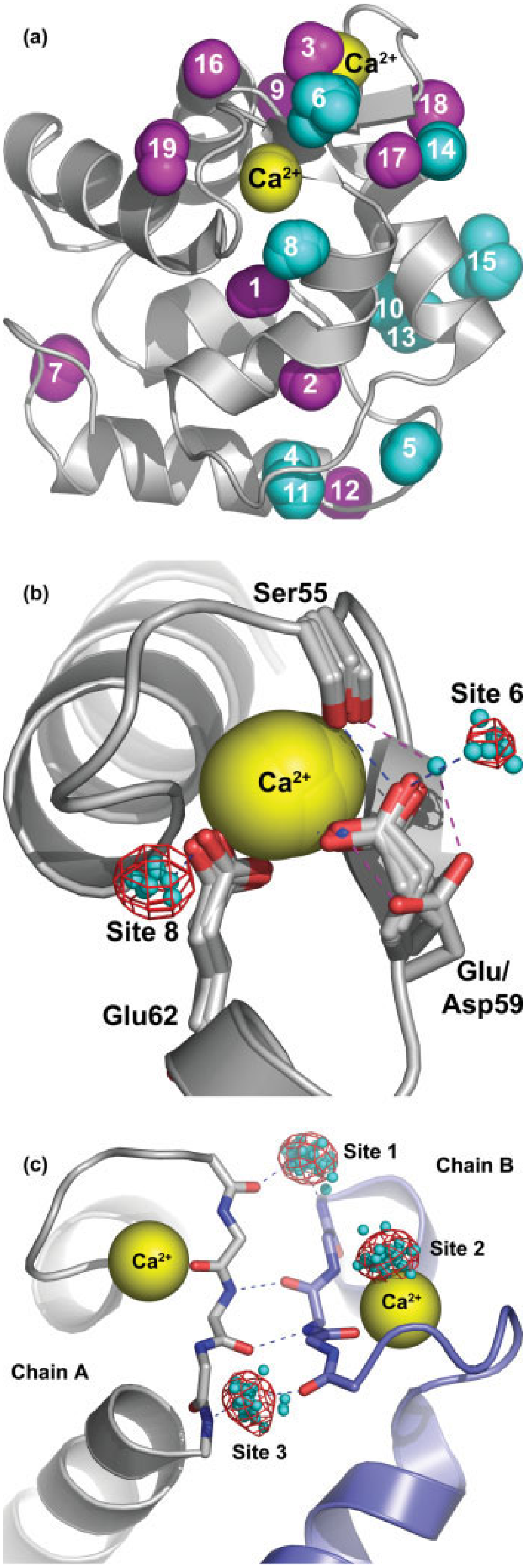
Serine Proteases

Several studies have been published on conserved solvent sites in serine proteases.[13,26,27,57,58] Our study discovered many conserved solvent sites in this protein, 10 of which had more than 80% conservation (Table VI). All of these had already been described in the literature. Perhaps this should not be surprising, because serine proteases are among the most studied proteins with respect to conserved solvent molecules. Two of these conserved sites (5 and 10) were found near the catalytic triad (Fig. 6). These two sites represent water molecules that are typically in hydrogen-bonding contact with each other. In some structures, they are replaced by three water molecules that perform similar roles. Site 5 bridges the amino nitrogen of either residue 101 or 102 to the N-terminus of a β-sheet [Fig. 6(b)]. Site 10 bridges the carbonyl oxygen of residue 99 to the hydroxyl of Ser214. Note that Ser214 interacts with the aspartate and histidine of the catalytic triad.

The two sites with more than 90% conservation are located near residues that have high sequence conservation. Site 1 is within 5 Å of a 100% conserved glycine residue (Gly211 in 1FY4). The complete conservation of this residue was not apparent from an initial ClustalW alignment, but was found from a manual inspection of superimposed residues. Site 2 is within 4 Å of a highly conserved tryptophan residue (89% sequence identity). When present, this tryptophan is always flanked by glycine residues. In two cases (1G2L[59] and 1GJ7[60]) a phenylalanine replaces the tryptophan, but it is still flanked by glycines. In two other cases, this motif is completely missing from the structure (1CQQ[61] and 1MBM[62]).

Effects of B-Values

Atomic B-values appear in the Fourier transform equations used for map calculation, so we examined the effects of these parameters on the results. The data



shown above were obtained by setting all B-values to 20 Å^2. We also performed the calculations using the B-values supplied in the PDB entries. As expected, the peak heights and locations differed slightly from those obtained with B = 20 Å^2, however; the rank ordering of sites, percent conservation values, and representative interactions were identical to those listed in Tables I–VI. Thus, our method seems to be insensitive to the B-factor model used. We suspect that this is attributable to the fact that proximity of water molecules to peaks in the pseudo-electron density map is only one of two criteria used for determining conservation, the other being interactions with the protein.

Influence of Crystal Packing

Protein crystals typically have solvent content near 50%[63] and so it is not surprising that water molecules often mediate protein–protein interactions in crystals. We inspected the crystal-packing environments around the conserved water molecules listed in Tables I–VI using Coot[64] to determine how many engage in crystal contacts. In the representative structures, only 5 of the 48 conserved water molecules formed an interaction within 3.5 Å of a symmetry-related protein: site 3 of cytochrome *c* (1YCC) and sites 8, 12, 16, and 18 of parvalbumin (2PVB). We then inspected these five sites in other structures and found several examples in which water molecules in these sites were free of crystal contacts. For example, site 3 of the cytochrome *c* family does not form crystal contacts in 1CCR, 1QN2, 1HRO, and 1WEJ. Parvalbumin sites 8, 12, 16, and 18 do not form crystal contacts in 1RRO or in 1PVA (chain B).

These results show that highly conserved sites are overwhelmingly free of crystal-packing contacts. This result is consistent with the observation that water molecules listed in Tables I–VI form an average of 2.1 hydrogen bonds with their associated proteins. Thus, highly conserved water molecules are tightly hydrogen bonded to the protein and have limited hydrogen-bond capacity left over for forming crystal contacts. Further-

Fig. 4. Structurally conserved water molecules in parvalbumins (a, b) and in the EF-hand superfamily (c). **a:** Pike β-parvalbumin (2PVB) with water molecules from the superposition. Percent conservation values for these sites can be found in Table IV. Cyan spheres denote novel solvent sites discovered by our methodology and magenta spheres denote sites identified by our method that have also been previously reported in the literature. **b:** Close-up view of conserved solvent sites 6 and 8 of parvalbumins. Water molecules belonging to these sites are drawn as small cyan spheres and the water-density map is contoured at 10σ. The water molecule that serves as the −x ligand in rat β-parvalbumin (1RRO) is shown as a small blue sphere. The water molecule of site 6 that is shown furthest to the left is from 1RRO. Magenta dotted lines refer to interactions found in 1RRO. **c:** Close-up view of the top three conserved solvent sites of the EF-hand superfamily. EF-hand superfamily sites 1, 2, and 3 correspond to sites 16, 3, and 17 of the parvalbumin family, respectively (Table IV). Water molecules belonging to these sites are drawn as small cyan spheres and the water-density map is contoured at 10σ. The protein shown is timothy grass (*Phleum pratense*) pollen allergen Phl p 7 (1K9U). Unlike other EF-hand proteins, the structurally adjacent calcium-binding sites are from different polypeptide chains. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

TABLE IV. Selected Conserved Solvent Sites of Parvalbumins

				Representative interactions[a]		
Site	Novel site?	Conservation (%)	Density peak (σ)	Backbone	Side-chain	Water ID[a]
1		100	30	O-Ile50 O-Glu62		204
2		100	30	O-Lys64 N-Leu67 O-Arg75		202
3		100	28		OD1-Asp94 OD2-Asp94 OE1-Glu101	201
4	√	100	27	O-Leu15		211
5	√	100	24	O-Ala76		284
6	√	100	15		OE2-Glu59	295
7		86	26	N-Lys7 O-Val33	OD2-Asp10	216
8	√	86	26		OE2-Glu62	212
9		86	25	O-Gly89 O-Glu101		203
10	√	86	25	O-Glu81		246
11	√	86	24	N-Arg75		230
12		86	23	O-Ala17 N-Ala20		218
13	√	86	20	O-Phe24		261
14	√	86	19	O-Gly95		257
15	√	86	17	O-Ala80		205
16		71	21	O-Gly56 N-Ile99		208[b]
17		71	21	N-Glu60 O-Gly95		224
18		71	17		OD2asp90	210
19		71	16		OD2asp51	221

[a] Representative interactions and water ID for structure with PDB code 2PVB.
[b] Symmetry-mate of water 208.

TABLE V. Selected Conserved Solvent Sites of Phospholipase A_2

				Representative interactions[a]		
Site	Novel site?	Conservation (%)	Density peak (σ)	Backbone	Side-chain	Water ID[a]
1	√	100	45	O-Cys1044		2005
2	√	100	43	O-Cys1098		2002
3		100	39	N-Ser1001 O-Ser1068	OD2-Asp1099	2006
4	√	92	39	N-Thr1041	OD2-Asp1039 OG1-Thr1041 OG1-Thr1112	2003
5	√	92	33	O-Ala1101		2173
6	√	92	38	N-Cys1084	OE2-Glu1097	2016
7	√	83	36	O-Leu1106	OH-Tyr1022 OG1-Thr1041	2001
8	√	83	27	O-Lys1049		2083
9		67	26		ND1-His1048 OD1-Asp49[b]	2009

[a] Except where noted, representative interactions and water ID for structure with PDB code 1MC2.
[b] Interaction seen in less than half the structures. Residue ID shown is from 1G4I.

more, we believe that our approach of comparing many structures with different space groups and unit cell parameters tends to filter out water molecules that are artificially stabilized by crystal-packing forces in any one particular lattice.

Effects of X-Ray Diffraction Data Collection Temperature

Given the preponderance of cryogenic structures in our study and the possibility for bias toward cryogenic data,

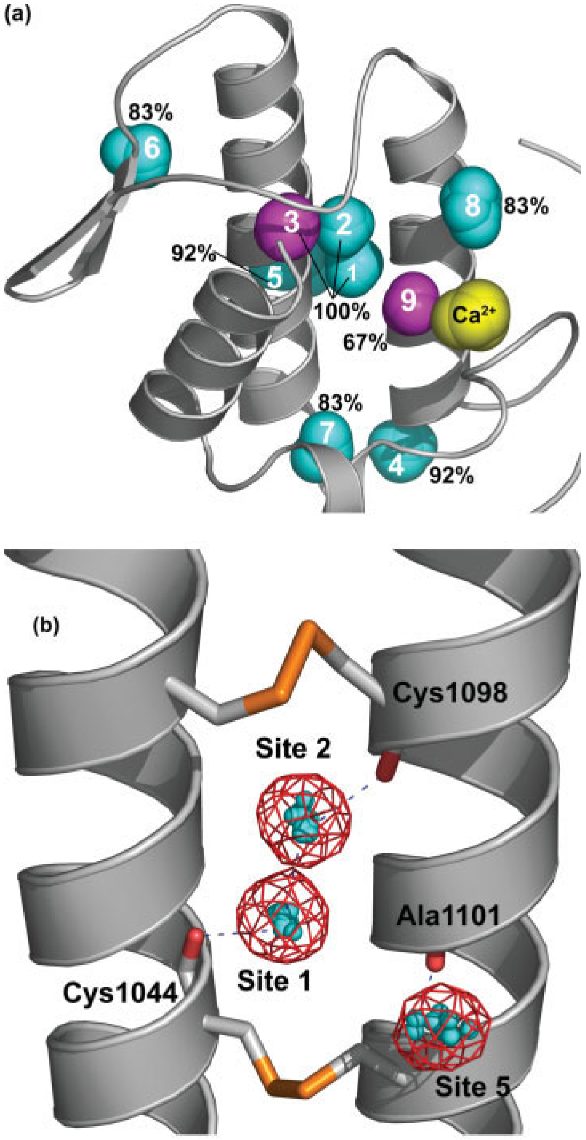


Fig. 5. Structurally conserved water molecules in the phospholipase A_2 family. **a:** Phospholipase A_2 from *Agkistrodon acutus* venom (1MC2) with water molecules from the superposition. Cyan spheres denote novel solvent sites discovered by our methodology and magenta spheres denote sites identified by our method that have also been previously reported in the literature. The percentage conservation values of the sites are indicated. **b:** Close-up view of sites 1, 2, and 5 (Table V). Sites 1 and 2 bridge backbone carbonyls of conserved disulfide-linked Cys. Site 5 forms a hydrogen with Ala1101 and makes van der Waals contact with one of the conserved Cys. Water molecules belonging to these sites are drawn as small cyan spheres and the water-density map is contoured at 10σ. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

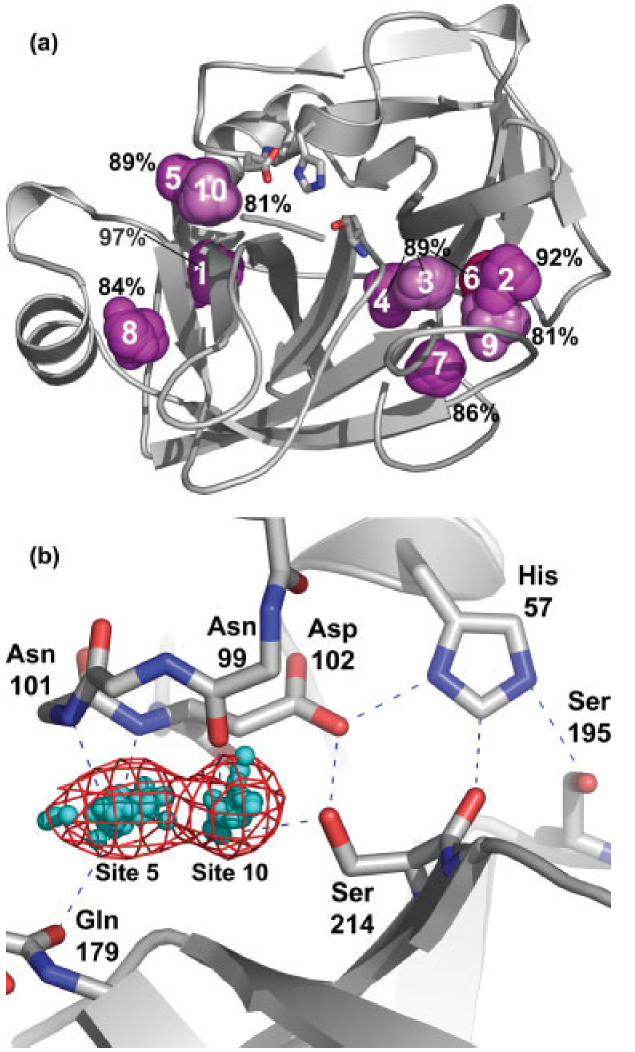


Fig. 6. Structurally conserved water molecules in serine proteases. **a:** *Fusarium oxysporum* trypsin (1FY4) with water molecules from the superposition. Residues of the catalytic triad are depicted as sticks. Magenta spheres denote solvent sites identified by our method that have also been previously reported in the literature. The percentage conservation values of the sites are indicated. **b:** Close-up view of sites 5 and 10 (Table VI), which are near the catalytic triad (Asp102, His57, Ser195). Water molecules belonging to these sites are drawn as small cyan spheres and the water-density map is contoured at 10σ. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

we asked whether water molecules identified by our analysis were present in RT structures. There are five structures in our cytochrome *c* data set with annotated data collection temperature near RT (1CCR, 1CO6, 1JDL, 1CXC, 1QN2). Sites 2 and 3 are present in all 5 RT structures, whereas site 1 appears in 4/5 RT structures and site 4 is present in 3/5 RT structures. However, the cryogenic structure 1I8O lacks site 3, and the cryogenic structure 1QL3 lacks site 4. Thus, it can be seen that the absence of these sites does not strictly correlate with data collection temperature.

1FDQ is the only annotated RT structure in our fatty-acid binding protein data set. Sites 1 and 3 are present in 1FDQ, but site 2 is absent. We note that one of the cryogenic structures (1B56) also lacks this water molecule.

The lactate/malate dehydrogenase data set includes two RT structures (1BDM, 1HYE). Site 1 is present in both RT structures, whereas sites 2 and 3 are missing from 1HYE and 1BDM, respectively. But, there are also cryogenic structures that lack sites 2 and 3.

Although none of the seven parvalbumin PDB entries had data collection temperature annotations near RT, the

TABLE VI. Selected Conserved Solvent Sites of Serine Proteases

Site	Novel site?	Conservation (%)	Density peak (σ)	Representative interactions[a] Backbone	Side-chain	Water ID[a]
1		97	40	O-Ile210 N-Gly232		1020
2		92	37	O-Ile30 O-Arg-66[b] N-Ser70		1011
3		89	43	N-Trp141 O-Gly193		1003
4		89	42	O-Ala139 O-Asp194		1002
5		89	36	N-Asn101 N-Asp102 O-Gln179[c]	OG1-Thr229[c,d]	1005[e]
6		89	35	O-Pro28 O-Ile30 N-Gly69		1024
7		86	42	N-Ile16 O-Gly140 O-Gly142		1014
8		84	26	O-Val163 N-Cys182		1018
9		81	34	O-Glu70[b]		67[b]
10		81	33	O-Asn99	OGser214	1008

[a] Except where noted, representative interactions and water ID for structure with PDB code 1FY4.
[b] From 1HJ8, because it more accurately represents the family in this instance.
[c] Found in less than half of the structures.
[d] From 1GVK, because it more accurately represents the family in this instance.
[e] This site has two waters in some structures.

primary literature sources for 1RRO,[65] 4CPV,[66] and 5PAL[42] suggested that data for these structures were obtained at RT. All three RT structures have sites 1–9, 12–16, and 19. Two of these three structures have water molecules at sites 10, 11, 17, and 18. Thus, all the conserved solvent sites listed in Table IV are represented in RT structures.

Four phospholipase A_2 structures were annotated as RT structures (1JIA, 1M8R, 1QLL, 1VIP). All four structures have water molecules at sites 1–6 and 8. Three RT structures (1JIA, 1M8R, and 1QLL) also have water molecules at sites 7 and 9. However, we note that some cryogenic structures lack water molecules that are found in the RT structures. For example, the cryogenic structure 1G4I lacks a water molecule at site 5 and 1LE6 lacks a water molecule at site 7.

Finally, there are six structures in the serine protease data set with annotated data collection temperature near RT (1A0J, 1CGH, 1EQ9, 1GJ7, 1GL1, 1H4W). Sites 1 and 3–8 (Table VI) are present in all six RT structures, whereas sites 2 and 10 are present in 5 of 6 RT structures. Site 9 is found in half of the RT structures. As with the other families, we found that some cryogenic structures lack water molecules that are found in the RT structures.

To summarize, in every family studied, the highly conserved water molecules were present to approximately the same extent in both cryogenic and RT structures. Moreover, in some cases, a cryogenic structure lacked a water molecule that was present in RT structures. These results show no correlation between conservation and data collection temperature for highly conserved water molecules (conservation ~70–100%).

DISCUSSION

Advantages and Limitations of the Method

Structural conservation of solvent sites in six protein families was studied using a method that represents the distribution of water in superimposed structures as a pseudo-electron density map. The method described herein has advantages over other methods that have been used to identify conserved solvent sites. For example, manual inspection using a graphics program has been the most frequently used approach. The main advantages of our computational approach are speed and objectivity. The only a priori knowledge required is a set of superimposed structures. No previous knowledge of conserved structural elements is required, such as the location of the active site or binding site. Thus, our method provides a powerful discovery tool. Manual analysis tends to focus on water molecules near the active sites or binding sites of proteins. Our method, however, is unbiased. It will discover conserved solvent sites in any part of the structure.

Other computational methods have been developed for analysis of protein-bound water. For example, Nakasako's[67] FESTKOP program also uses a density-based approach. A strength of FESTKOP is that it classifies waters into three divisions: inside the solvent accessible surface, crystal contact, and outside the solvent accessible surface.

TABLE VII. Pairwise Amino Acid Sequence Identity Statistics for the Data Sets Used in This Study†

Family	n	Mean	Median	Minimum	Maximum	SD	$\binom{n}{2}$
Cytochrome *c*	14	44	42	31	84	9	91
Fatty-acid binding protein	11	39	38	15	74	13	55
Lactate/malate dehydrogenase	14	30	28	12	77	11	91
Rossmann-fold dinucleotide binding	126	14	13	1	88	6	7,875
Parvalbumins	7	59	55	44	89	14	21
EF-hand superfamily	38	20	18	1	86	11	703
Phospholipase A_2	12	49	48	33	82	11	66
Serine proteases	37	35	35	9	87	11	666

†n = number of nonredundant structures in the data set. The last column is the number of one-on-one comparisons.

These divisions facilitate display and analysis of hydration structures and we note that this classification scheme could be incorporated into our method. FESTKOP has been used to study the solvent structure in glutamate dehydrogenase,[68] lysozyme,[69,70] killer toxin from *Pichia farinosa*,[71] and different crystal forms of bovine β-trypsin.[67]

Sanschagrin and Kuhn[13] have described a semiautomated method for identifying conserved solvent sites based on complete linkage cluster analysis. Using this method, they found conserved water molecules among a set of thrombin structures (minimum pairwise sequence identity >96%) and among a set of trypsin structures (100% sequence identity). The results of each of these two analyses were compared to find solvent sites conserved between thrombin and trypsin. However, complete linkage cluster analysis was not directly used to compare thrombin and trypsin. Linkage cluster analysis has also been used to study conserved water in structures of *Bacillus stearothermophilus* alanine racemase that were crystallized with different ligands bound.[72]

In contrast to the aforementioned applications of FESTKOP and cluster analysis, we compared proteins having much higher structural and sequence diversity (Table VII). Because studies using linkage cluster analysis and FESTKOP focused on proteins with very high sequence identity, a direct comparison with our method is not possible at this time. We did, however, apply our method to the set of serine protease structures used in the Sanschagrin and Kuhn work, and we were able to reproduce their results with excellent quantitative agreement (results not shown).

The current implementation of our method does have limitations. First, it is dependent on the superposition method. We used a method that superimposes proteins by their global fit to other proteins. Because global fitting is based on similar folds, it is an inherently low resolution method. To improve results, fitting of domains and smaller substructures could be implemented. Iterative superpositioning could focus on protein regions near water-density peaks.

Another limitation of our method is that only water molecules seen in crystal structures are analyzed. Of these, only the most ordered solvent molecules will be consistently seen in multiple structures. In principle, there could be dynamic water molecules that are also conserved but that are not observed in crystal structures. To identify such mobile, yet conserved, solvent molecules will require additional techniques such as molecular dynamics simulation, which has been used previously to identify preferred hydration sites around proteins.[73,74]

Data Collection Temperature

The number of ordered water molecules that can be reliably modeled in protein crystal structures depends on resolution and quality of the diffraction data, which in turn, depend on the data collection temperature.[75] Typically, data collected at cryogenic temperatures yield more water molecules in the final structure than data collected at RT.[67] Thus, one might expect data collection temperature to be an important parameter in calculations of conserved solvent structure.

In principle, one could perform separate analyses for cryogenic and RT structures. In practice, many PDB entries do not have an annotated data collection temperature, and looking up the temperature in the literature is not feasible for the PDB-wide analyses that we envision in the future. Moreover, cryogenic data collection has been a mainstay of protein crystallography for more than 15 years[76] and the PDB is, and will continue to be, dominated by low-temperature structures. Thus, any bioinformatics approach, such as ours, that relies on large numbers of structures, is potentially biased toward cryogenic structures. Interestingly, we found no evidence for such bias in our results. As discussed above, we found that water molecules with ~70–100% conservation form an average of 2.1 hydrogen bonds with the protein. Thus, highly conserved water molecules are bound tightly to the protein, which might explain their presence in both low temperature and RT structures.

Other Applications

This study focused on conserved solvent in protein families, but our method could be applied to other problems involving protein-bound water. For example, conserved solvent sites could be analyzed within a single structure that has a repeating structural unit. In this context, our method was used recently to discover that the β-propeller of the Keap1 Kelch domain has multiple solvent sites that are structurally conserved among all six individual blades.[77] To our knowledge, this aspect of β-propeller structure had not been appreciated previously, which attests to the utility of our method.

Knowledge of conserved solvent could also be used to selectively model highly conserved water molecules in low/moderate-resolution crystal structures. Difference electron density maps for low/moderate-resolution structures will likely exhibit features corresponding to highly conserved solvent sites, although most crystallographers would be reluctant to model any solvent. For example, by using the Electron Density Server,[40] we observed an electron density peak in a 3.0 Å structure of human class IV alcohol dehydrogenase (1AGN)[78] that corresponds to the highly conserved water molecule of the Rossmann-fold dinucleotide-binding domain (peak 1, Table III). This feature is especially apparent in the D chain of 1AGN. Thus, a conserved water molecule could probably be built in this low-resolution structure, even if no other water molecules were included. Doing so in this case would have been substantiated by a 2.5 Å resolution structure (1D1S[79]) of the same protein deposited later in the PDB, which contained the structurally conserved water molecule modeled in all four chains in the asymmetric unit.

Protein structure modeling is another area of potential application. Because of advances in whole-genome sequencing, there are currently many more protein sequences known than three-dimensional protein structures, making protein structure modeling ever more important.[80,81] Homology modeling programs often treat solvent as a continuum fluid because of the computational and theoretical challenges associated with modeling explicit water molecules. Yet, bound water molecules are integral components of protein folds, and thus there are clear benefits to including water in protein models. For example, addition of conserved water to a homology model of rat submaxillary kallikrein greatly improved the Ramachandran statistics and led to a more accurate model.[58] We suggest that knowledge of highly conserved water molecules and the interactions that they form with the protein could provide a basis for selective incorporation of water into homology models. This application will be explored in the future.

Analogy to Multiple Amino Acid Sequence Alignment

Multiple amino acid sequence alignment is routinely used to identify functionally and/or structurally important residues that define protein families.[82] Proteins within a family exhibit various levels of sequence conservation throughout the polypeptide chain. Residues with 100% sequence identity represent one end of a continuum of amino acid sequence conservation that is often observed in multiple sequence alignments. Many other important residues exhibit high, but less than 100%, sequence identity because of subtle variations in structure, function, thermostability, substrate specificity, etc. At the other end of the spectrum, there will be some regions of a multiple sequence alignment that show no statistically significant amino acid sequence homology.

Our approach for understanding conservation of protein-bound water is analogous to multiple sequence alignment. We likewise observed a continuum of conservation for solvent sites for the six families studied here. As with amino acid sequence analysis, there are some solvent sites with 100% conservation (Tables II–V). Our interpretation is that these sites are especially important for structure, stability, or function, and that these absolutely conserved water molecules are defining features of the protein family. We also found several solvent sites with very high, but less than 100% conservation (Tables I–VI, 80–97% sites). We likewise interpret these sites as being important determinants of structure and/or function, although perhaps not as significant for defining the family as those with 100% conservation. There are also many solvent sites with low conservation (<50%), which likely have somewhat nonspecific roles in protein hydration.

There are several possible reasons why a particular protein might lack a certain highly conserved water molecule in our analysis, thus giving rise to sites with substantial, but less than 100% conservation (80–99%). First, the solvent model represented in the PDB file may be incomplete. For example, structure 1HYH does not have a water molecule corresponding to conserved site 3 of the lactate/malate dehydrogenase family. However, by using the Electron Density Server,[40] we observed a convincing feature in the $2F_o$-F_c map corresponding to this water molecule. Second, the binding of ligands could displace conserved waters. This is the case for phospholipase A_2 structures 1HN4 and 1KVO, in which an inhibitor displaces the catalytic water molecule. A third possibility is that our method failed to identify the water molecule in question. This could occur, for example, if the superposition is not optimal or if a water molecule is rejected during enforcement of distance cutoffs in the clustering or hydrogen-bonding steps (Fig. 7). The former issue could be addressed by iterative superposition, as discussed above. Cutoffs could be increased from the values used here, but this could potentially lead to false positives. Finally, by analogy to multiple sequence alignment, we suggest that subtle differences of protein structure and function within a family will result in highly conserved sites that truly have less than 100% conservation. These highly, but not absolutely, conserved sites thus potentially represent loci of structural variation that underlie differences in binding affinity, catalytic efficiency, substrate specificity, and protein stability.

Connections Between Solvent Structure Conservation and Sequence Conservation

We observed that solvent sites with more than 90% conservation are located near residues with very high sequence conservation. For example, site 1 of cytochrome *c* is near the heme-binding His residue and the conserved Gly-Pro motif of the Ω-loop [Fig. 1(e)]. These three residues are among only 12 that have 100% sequence identity in our cytochrome *c* data set. Cytochrome *c* site 2 forms a hydrogen bond to a heme propionate [Fig. 1(d)]. Fatty acid binding protein site 1 binds between conserved Gly and Phe residues, whereas site 2 contacts a conserved Phe near the N-terminus [Fig. 2(b)]. These three residues are among

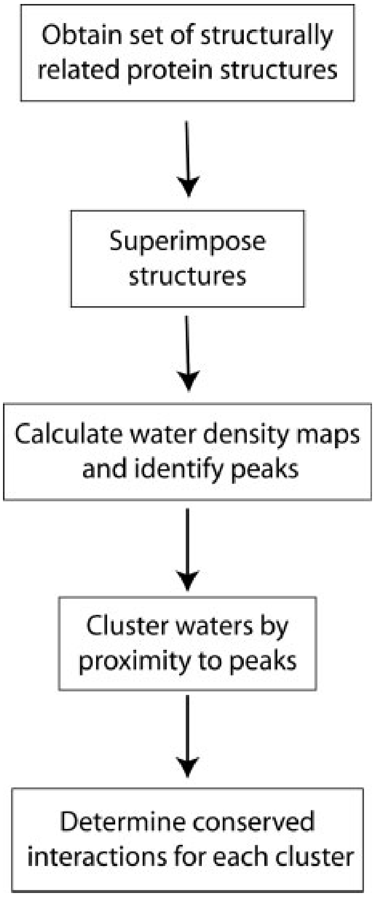

Obtain set of structurally related protein structures
Superimpose structures
Calculate water density maps and identify peaks
Cluster waters by proximity to peaks
Determine conserved interactions for each cluster

Fig. 7. Schematic diagram of the computational algorithm used in this study for discovery of structurally conserved water molecules.

only four residues that have 100% sequence identity in our fatty acid binding protein data set. Site 1 of lactate/malate dehydrogenase forms a hydrogen bond with the last Gly of the conserved Gly-rich loop of the Rossmann fold. The first and last Gly residues of this motif are two of only 11 residues that have 100% sequence identity in our data set. There are six solvent sites with 100% conservation in the parvalbumin data set. Each one is within 4.3 Å of a residue with 100% sequence identity except for site 6. In this case, the water molecule is next to an absolutely conserved carboxyl group of either Asp or Glu [Fig. 4(b)]. Phospholipase A2 sites 1, 2, 5, and 6 are next to completely conserved disulfide bonds (Cys1044—Cys1105, Cys1051—Cys1098, Cys1084—Cys1096 in 1MC2). Sites 3 and 4 form hydrogen bonds with the side-chains of conserved Asp1099 and Asp1039, respectively (1MC2 numbering). Finally, serine protease site 1 is within 5 Å of a 100% conserved glycine residue, site 2 binds next to a highly conserved G(W/F)G motif, and sites 5 and 10 are near the conserved catalytic triad.

Thus, there seems to be a spatial correlation between highly conserved solvent sites and highly conserved residues, at least for the structures studied here. Highly conserved residues tend to form conserved three-dimensional substructures that are important for folding, stability, binding, and catalysis, such as the Ω-loop of cytochrome *c*, the binding site of fatty-acid binding proteins, the pyrophosphate-binding loop of the Rossmann fold, EF-hand Ca^{2+}-binding loops of parvalbumin, disulfide bonds of phospholipase A_2, and the catalytic triad of serine protease. The observation that structurally conserved water molecules bind these substructures suggests that these water molecules might be particularly important for structure and function and adds a new dimension to our understanding of the conserved structural elements that define protein families.

CONCLUSION

For each protein family studied, our method successfully identified previously known conserved solvent sites, which validated our approach. Moreover, our method discovered 22 novel conserved solvent sites in protein families that have been intensely studied for decades and these discoveries suggested new structure/function relationships. All of the families studied had solvent sites with more than 90% conservation and these sites were invariably located near residues with very high sequence conservation. It is axiomatic in biology that patterns of conserved amino acid residues, as well as the conserved three-dimensional substructures that these residues form, are defining features of protein families.[82] By analogy, we hypothesize that highly conserved solvent sites, by virtue of their proximity to conserved residues, are also defining features of protein families. Because all the families studied here had highly conserved solvent sites (>90%), we further hypothesize that other protein families—perhaps all protein families—have highly conserved solvent sites that are diagnostic of the family. We plan to test and elaborate these ideas in the future by performing PDB-wide analyses of protein-bound water using the approach described here.

MATERIALS AND METHODS

Description of Method

Our strategy can be summarized as a pipeline of five steps, each consisting of one or more calculations (Fig. 7). First, a query protein structure was input to the SSM[83] service to retrieve a set of protein structures related by a common three-dimensional architecture (Fig. 7, step 1). The default parameters of SSM were used for this study. A nonredundant subset of the structures returned by SSM was identified by comparison to the PDB_SELECT list,[84] and structures with a high-resolution diffraction limit of 2.2 Å or better were retained. Coordinate files were retrieved from the PQS database. These files contain the full quaternary structure and all crystallographically related water molecules near the protein. The SSM service (also called MSDfold) and the PQS database can be found at the European Bioinformatics Institute's MSD Web site (http://www.ebi.ac.uk/msd/).

The nonredundant protein structures, including water molecules, were next superimposed using the transformation matrices and vectors obtained from SSM (Fig. 7, step 2). A locally written Fortran-90 program was used for this purpose but there are also several free programs such as

Moleman[85] that could be used to rotate and translate PDB coordinates. Another locally written Fortran-90 program was used to assemble all the water molecules from the superimposed structures into a single PDB-formatted file, renumber the water molecule residue numbers to avoid conflicts in CNS,[86] remove water molecules extremely distant from the surface of the reference structure, and calculate the dimensions of the resulting system.

An electron density map representing the distribution of water molecules in the data set was calculated using CNS (Fig. 7, step 3). B-factors of all water molecules were set to 20 Å^2 to avoid possible artifacts because of variations in B-factors from structure to structure, although this step is not necessary (see Results). A modified version of the CNS script generate_easy.inp was used to make molecular topology and CNS-formatted coordinate files for the superimposed water molecules. The number of water molecules used to calculate each density map ranged from 1,208 for parvalbumins to 38,845 for the Rossmann fold data set. Theoretical structure factors were calculated with the CNS script model_fcalc.inp using a P1 lattice with unit cell angles of 90° and unit cell lengths 5 Å larger than the dimensions of the system of waters. We emphasize that this calculation includes water molecules only. The CNS script make_cv.inp was used to add the test flag required by subsequent CNS scripts. The resulting calculated structure factors were input to the CNS script model_map.inp to calculate a pseudo-electron density map ($|F_c|$, ϕ_c) and generate a list of peak locations and peak heights. We emphasize that this calculation includes water molecules only. Each peak represents a potential conserved solvent site of the protein family.

The water molecules in the data set were next organized into clusters (Fig. 7, step 4) based on their proximity to the peaks of the map using a 2.0 Å cutoff. A locally written C# program was used for this calculation. Finally, the noncovalent interactions formed by each water molecule in its respective structure were tabulated using the C# program (Fig. 7, step 5). For this study, two atoms were defined as interacting if they were separated by no more than 3.2 Å. This criterion focuses on hydrogen bonds, but it could easily be modified to include van der Waals interactions, or it could be made more restrictive by including an angle cutoff in hydrogen-bond determination.

Water molecules in different structures are considered to be structurally conserved if (1) they belong to the same peak cluster, and (2) they have at least one equivalent interaction with the protein in common. We define percent conservation as the percentage of structures within a data set that contain structurally equivalent water molecules within a given cluster. This value is analogous to percent conservation of an amino acid residue at a specific position within a multiple sequence alignment.

Although many conserved solvent sites exist, we have limited ourselves to listing only those that have the highest conservation (e.g., cytochrome *c* sites 1–3, Table I) or, in several cases, those that have interesting features or that have received considerable attention in the literature (e.g., cytochrome *c* site 4, Table I).

Programs used for molecular visualization include O,[87] Protein Explorer,[88] PyMOL,[89] and Coot.[64] The CNS scripts, FORTRAN programs, and C# program used for this study are available from the authors upon request.

Data Sets Used in This Study

The method was tested on six protein families for which conserved solvent sites have been previously identified: cytochrome *c*, fatty-acid binding protein, lactate/malate dehydrogenase, parvalbumin, phospholipase A$_2$, and serine proteases. For each data set containing *n* structures, $n(n-1)/2$ pairwise sequence identities were obtained using SSM.[83] In the case of serine proteases, 10 structures were removed from the data set because they contained more than 90% sequence identity. The resulting sequence comparison statistics for each of these data sets is given in Table VII.

The representative structure of the cytochrome *c* family that was input to SSM was 1YCC,[30] and the resulting nonredundant data set consisted of 14 structures: 1CCR, 1CO6, 1COT, a revised 1CXC structure,[90] 1HRO, 1I8O, 1JDL, 1QL3, 1QN2, 1WEJ, 1YCC, 1YTC, 3C2C, 5CYT. For the fatty-acid binding protein family, the representative structure was 1HMT[91] and the nonredundant data set included 11 structures: 1B56, 1CBS, 1CRB, 1FDQ, 1FTP, 1HMT, 1KQW, 1LID, 1LPJ, 1MDC, 1OPB. For lactate/malate dehydrogenases, the representative structure was 1LDG[92] and the resulting data set consisted of 14 structures: 1A5Z, 1BDM, 1GUY, 1GUZ, 1HYE, 1HYH, 1I0Z, 1LDG, 1LLD, 1MLD, 1O6Z, 2CMD, 6LDH, 9LDT. The representative of the parvalbumin family was 2PVB[41] and the data set consisted of seven structures: 1A75, 1BU3, 1PVA, 1RRO, 2PVB, 4CPV, 5PAL. The phospholipase A$_2$ data set included 12 structures, resulting from using 1MC2[93] as the representative structure: 1FV0, 1G4I, 1HN4, 1JIA, 1JLT, 1KVO, 1LE6, 1M8R, 1MC2, 1QLL, 1VAP, 1VIP. The serine protease data set included 37 structures, with 1FY4[94] as the reference: 1A0J, 1A7S, 1BIO, 1CGH, 1CQQ, 1DDJ, 1EAX, 1ELT, 1EQ9, 1F7Z, 1FI8, 1FIW, 1FUJ, 1FXY, 1FY4, 1G2L, 1GJ7, 1GL1, 1GVK, 1H4W, 1H8D, 1HJ8, 1HJ9, 1IAU, 1KLI, 1LO6, 1MBM, 1MCT, 1NN6, 1NPM, 1PPF, 1SGT, 1TON, 1TRN, 1UCY, 2HLC, 3RP2.

Analysis of conserved solvent was also performed for Rossmann-fold proteins and for the EF-hand superfamily. For these calculations, the PDB Web site[29] was used to identify structurally similar proteins. Using new selection features of this site, PDB files were chosen such that they (1) were within the same SCOP classification, (2) were crystal structures with 2.2 Å or better resolution, and (3) were not more than 90% identical in sequence (using the CD-HIT algorithm).[95] Because of the large size of this data set, we used the map and peak picking utilities of CCP4[96] rather than CNS.

The EF-hand data set was based on the SCOP superfamily classification “EF-hand” (SCOP code: a.39.1). The EF-hand superfamily included 38 structures: 1ALV, 1AUI, 1BU3, 1CDP, 1DTL, 1E8A, 1EG3, 1EXR, 1G4Y, 1G8I, 1GGZ, 1IG5, 1IRJ, 1J55, 1JF0, 1K8U, 1K94, 1K9U, 1KSO, 1M45, 1MHO, 1MR8, 1NCX, 1OE9, 1OMR, 1PSR, 1PVA, 1QV1, 1RRO, 1RWY, 1S6C, 1SRA, 1UHN, 1WDC, 2CBL, 2PVB, 2SCP, 5PAL.

The Rossmann-fold protein data set was based on the SCOP fold classification "NAD(P)-binding Rossmann-fold domains" (SCOP code: c.2). The nonredundant data set (n = 126) of structures within the SCOP fold classification "NAD(P)-binding Rossmann-fold domains" included the following: 1A4I, 1A5Z, 1B16, 1B7G, 1B8P, 1BDB, 1BDM, 1BG6, 1BGV, 1BXK, 1CDO, 1CF2, 1CYD, 1D7O, 1DLJ, 1DPG, 1DSS, 1DXY, 1E3I, 1E5Q, 1E6U, 1E6W, 1E7W, 1EK6, 1EQ2, 1EUD, 1EVY, 1F06, 1F0Y, 1FJH, 1FMC, 1G0O, 1GAD, 1GD1, 1GEE, 1GEG, 1GPJ, 1GR0, 1GU7, 1GUZ, 1GY8, 1H2B, 1H5Q, 1H6D, 1HDO, 1HEU, 1HT0, 1HXH, 1HYE, 1I0Z, 1I24, 1I36, 1IUK, 1IY8, 1J3V, 1J4A, 1J5P, 1JA9, 1JAY, 1JQB, 1JTV, 1JVB, 1K3T, 1K6X, 1KEW, 1KOL, 1KS9, 1L7D, 1LC0, 1LDM, 1LI4, 1LJ8, 1LLD, 1LUA, 1M6H, 1MB4, 1MG5, 1MLD, 1MV8, 1MX3, 1N2S, 1N7H, 1NP3, 1NPY, 1NVM, 1NWH, 1NXQ, 1NYT, 1O0S, 1O6Z, 1OAA, 1OBB, 1OBF, 1OC2, 1OI7, 1ORR, 1P0F, 1P1J, 1P77, 1PJ3, 1PJC, 1PL8, 1PR9, 1PX0, 1PZG, 1Q7B, 1QMG, 1QSG, 1R6D, 1RKX, 1RPN, 1T2A, 1T2D, 1UAY, 1UDC, 1UR5, 1UUF, 1VI2, 1VJ0, 1VJ1, 1VJP, 2AE2, 2CMD, 2NAC, 2PGD, 9LDT.

ACKNOWLEDGMENTS

Support for C.A.B. was provided, in part, by the Interdisciplinary Plant Group and the Genetics Area Program of the University of Missouri-Columbia. The authors thank Dr. Herbert Axelrod for providing revised coordinates of 1CXC, and Drs. Eugene Krissinel and Kim Henrick for help with SSM and MSD. C.A.B. thanks Matthew L. Klein for programming assistance. The authors acknowledge the many crystallographers whose structures were used in this analysis. They regret not being able to provide a bibliographic citation for each structure listed in Materials and Methods because of space limitations.

REFERENCES

1. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res 2001;29:2860–2874.
2. Janin J. Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. Structure 1999;7:R277–279.
3. Bottoms CA, Smith PE, Tanner JJ. A structurally conserved water molecule in Rossmann dinucleotide-binding domains. Protein Sci 2002;11:2125–2137.
4. Babor M, Sobolev V, Edelman M. Conserved positions for ribose recognition: importance of water bridging interactions among ATP, ADP and FAD-protein complexes. J Mol Biol 2002;323:523–532.
5. Powers RA, Shoichet BK. Structure-based approach for binding site identification on AmpC beta-lactamase. J Med Chem 2002;45:3222–3234.
6. Ni H, Sotriffer CA, McCammon JA. Ordered water and ligand mobility in the HIV-1 integrase-5CITEP complex: a molecular dynamics study. J Med Chem 2001;44:3043–3047.
7. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 2002;46:34–40.
8. Rarey M, Kramer B, Lengauer T. The particle concept: placing discrete water molecules during protein-ligand docking predictions. Proteins 1999;34:17–28.
9. Johnson LN, De Moliner E, Brown NR, et al. Structural studies with inhibitors of the cell cycle regulatory kinase cyclin-dependent protein kinase 2. Pharmacol Ther 2002;93:113–124.
10. Rao MS, Olson AJ. Modelling of factor Xa-inhibitor complexes: a computational flexible docking approach. Proteins 1999;34:173–183.
11. Engh RA, Brandstetter H, Sucher G, et al. Enzyme flexibility, solvent and 'weak' interactions characterize thrombin-ligand interactions: implications for drug design. Structure 1996;4:1353–1362.
12. Katz BA, Elrod K, Luong C, et al. A novel serine protease inhibition motif involving a multi-centered short hydrogen bonding network at the active site. J Mol Biol 2001;307:1451–1486.
13. Sanschagrin PC, Kuhn LA. Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. Protein Sci 1998;7:2054–2064.
14. Vogt J, Perozzo R, Pautsch A, et al. Nucleoside binding site of herpes simplex type 1 thymidine kinase analyzed by X-ray crystallography. Proteins 2000;41:545–553.
15. Rutenber EE, Stroud RM. Binding of the anticancer drug ZD1694 to *E. coli* thymidylate synthase: assessing specificity and affinity. Structure 1996;4:1317–1324.
16. Sage CR, Rutenber EE, Stout TJ, Stroud RM. An essential role for water in an enzyme reaction mechanism: the crystal structure of the thymidylate synthase mutant E58Q. Biochemistry 1996;35:16270–16281.
17. Wang T, Wade RC. Comparative binding energy (COMBINE) analysis of influenza neuraminidase-inhibitor complexes. J Med Chem 2001;44:961–971.
18. Mancera RL. De novo ligand design with explicit water molecules: an application to bacterial neuraminidase. J Comput Aided Mol Des 2002;16:479–499.
19. Minke WE, Diller DJ, Hol WG, Verlinde CL. The role of waters in docking strategies with incremental flexibility for carbohydrate derivatives: heat-labile enterotoxin, a multivalent test case. J Med Chem 1999;42:1778–1788.
20. Faerman CH, Karplus PA. Consensus preferred hydration sites in six FKBP12-drug complexes. Proteins 1995;23:1–11.
21. Likic VA, Juranic N, Macura S, Prendergast FG. A "structural" water molecule in the family of fatty acid binding proteins. Protein Sci 2000;9:497–504.
22. Berghuis AM, Guillemette JG, McLendon G, Sherman F, Smith M, Brayer GD. The role of a conserved internal water molecule and its associated hydrogen bond network in cytochrome *c*. J Mol Biol 1994;236:786–799.
23. Loris R, Stas PP, Wyns L. Conserved waters in legume lectin crystal structures. The importance of bound water for the sequence-structure relationship within the legume lectin family. J Biol Chem 1994;269:26722–26733.
24. Kumar A, Sekharudu C, Ramakrishnan B, et al. Structure and function of the catalytic site mutant Asp 99 Asn of phospholipase A_2: absence of the conserved structural water. Protein Sci 1994;3:2082–2088.
25. Loris R, Langhorst U, De Vos S, et al. Conserved water molecules in a large family of microbial ribonucleases. Proteins 1999;36:117–134.
26. Krem MM, Di Cera E. Conserved water molecules in the specificity pocket of serine proteases and the molecular mechanism of Na^+ binding. Proteins 1998;30:34–42.
27. Sreenivasan U, Axelsen PH. Buried water in homologous serine proteases. Biochemistry 1992;31:12785–12791.
28. Bottoms CA, Schuermann JP, Agah S, Henzl MT, Tanner JJ. Crystal structure of rat α-parvalbumin at 1.05 Å resolution. Protein Sci 2004;13:1724–1734.
29. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank (http://www.rcsb.org/). Nucleic Acids Res 2000;28:235–242.
30. Louie GV, Brayer GD. High-resolution refinement of yeast iso-1-cytochrome *c* and comparisons with other eukaryotic cytochromes *c*. J Mol Biol 1990;214:527.
31. Lett CM, Berghuis AM, Frey HE, Lepock JR, Guillemette JG. The role of a conserved water molecule in the redox-dependent thermal stability of iso-1-cytochrome *c*. J Biol Chem 1996;271:29088–29093.
32. Sogabe S, Miki K. Refined crystal structure of ferrocytochrome c_2 from *Rhodopseudomonas viridis* at 1.6 Å resolution. J Mol Biol 1995;252:235–247.

33. Benning MM, Meyer TE, Holden HM. Molecular structure of a high potential cytochrome c_2 isolated from *Rhodopila globiformis*. Arch Biochem Biophys 1996;333:338–348.
34. Geremia S, Garau G, Vaccari L, et al. Cleavage of the iron-methionine bond in c-type cytochromes: crystal structure of oxidized and reduced cytochrome c_2 from *Rhodopseudomonas palustris* and its ammonia complex. Protein Sci 2002;11:6–17.
35. Benning MM, Meyer TE, Holden HM. X-Ray structure of the cytochrome c_2 isolated from *Paracoccus denitrificans* refined to 1.7-Å resolution. Arch Biochem Biophys 1994;310:460–466.
36. Likic VA, Prendergast FG. Dynamics of internal water in fatty acid binding protein: computer simulations and comparison with experiments. Proteins 2001;43:65–72.
37. LaLonde JM, Bernlohr DA, Banaszak LJ. The up-and-down β-barrel proteins. FASEB J 1994;8:1240–1247.
38. Dengler U, Niefind K, Kiess M, Schomburg D. Crystal structure of a ternary complex of D-2-hydroxyisocaproate dehydrogenase from *Lactobacillus casei*, NAD$^+$ and 2-oxoisocaproate at 1.9 Å resolution. J Mol Biol 1997;267:640–660.
39. Niefind K, Hecht HJ, Schomburg D. Crystal structure of L-2-hydroxyisocaproate dehydrogenase from *Lactobacillus confusus* at 2.2 Å resolution. An example of strong asymmetry between subunits. J Mol Biol 1995;251:256–281.
40. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA. The Uppsala Electron-Density Server. Acta Crystallogr D Biol Crystallogr 2004;60:2240–2249.
41. Declercq JP, Evrard C, Lamzin V, Parello J. Crystal structure of the EF-hand parvalbumin at atomic resolution (0.91 Å) and at low temperature (100 K). Evidence for conformational multistates within the hydrophobic core. Protein Sci 1999;8:2194–2204.
42. Roquet F, Declercq JP, Tinant B, Rambaud J, Parello J. Crystal structure of the unique parvalbumin component from muscle of the leopard shark (*Triakis semifasciata*). The first X-ray study of an α-parvalbumin. J Mol Biol 1992;223:705–720.
43. Strynadka NC, James MN. Crystal structures of the helix-loop-helix calcium-binding proteins. Annu Rev Biochem 1989;58:951–998.
44. Kretsinger RH, Nockolds CE. Carp muscle calcium-binding protein. II. Structure determination and general description. J Biol Chem 1973;248:3313–3326.
45. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
46. Bourne Y, Dannenberg J, Pollmann V, Marchot P, Pongs O. Immunocytochemical localization and crystal structure of human frequenin (neuronal calcium sensor 1). J Biol Chem 2001;276: 11949–11955.
47. Hohenester E, Maurer P, Hohenadl C, Timpl R, Jansonius JN, Engel J. Structure of a novel extracellular Ca^{2+}-binding module in BM-40. Nat Struct Biol 1996;3:67–73.
48. Verdino P, Westritschnig K, Valenta R, Keller W. The cross-reactive calcium-binding pollen allergen, Phl p 7, reveals a novel dimer assembly. EMBO J 2002;21:5007–5016.
49. Moroz OV, Antson AA, Murshudov GN, et al. The three-dimensional structure of human S100A12. Acta Crystallogr D Biol Crystallogr 2001;57:20–29.
50. Kilby PM, Van Eldik LJ, Roberts GC. The solution structure of the bovine S100B protein dimer in the calcium-free state. Structure 1996;4:1041–1052.
51. Itou H, Yao M, Fujita I, et al. The crystal structure of human MRP14 (S100A9), a Ca^{2+}-dependent regulator protein in inflammatory process. J Mol Biol 2002;316:265–276.
52. Vijay-Kumar S, Cook WJ. Structure of a sarcoplasmic calcium-binding protein from *Nereis diversicolor* refined at 2.0 Å resolution. J Mol Biol 1992;224:413–426.
53. Brunie S, Bolin J, Gewirth D, Sigler PB. The refined crystal structure of dimeric phospholipase A$_2$ at 2.5 Å. Access to a shielded catalytic center. J Biol Chem 1985;260:9742–9749.
54. Dijkstra BW, Drenth J, Kalk KH. Active site and catalytic mechanism of phospholipase A$_2$. Nature 1981;289:604–606.
55. Epstein T, Yu B, Pan Y, et al. The basis for k*$_{cat}$ impairment in prophospholipase A$_2$ from the anion-assisted dimer structure. Biochemistry 2001;40:11411–11422.
56. Cha SS, Lee D, Adams J, et al. High-resolution X-ray crystallography reveals precise binding interactions between human nonpancreatic secreted phospholipase A$_2$ and a highly potent inhibitor (FPL67047XX). J Med Chem 1996;39:3878–3881.
57. Di Cera E, Guinto ER, Vindigni A, et al. The Na$^+$ binding site of thrombin. J Biol Chem 1995;270:22089–22092.
58. Henriques EF, Ramos MJ, Reynolds CA. Inclusion of conserved buried water molecules in the model structure of rat submaxillary kallikrein. J Comput Aided Mol Des 1997;11:547–556.
59. Nar H, Bauer M, Schmid A, et al. Structural basis for inhibition promiscuity of dual specific thrombin and factor Xa blood coagulation inhibitors. Structure (Camb) 2001;9:29–37.
60. Katz BA, Sprengeler PA, Luong C, et al. Engineering inhibitors highly selective for the S1 sites of Ser190 trypsin-like serine protease drug targets. Chem Biol 2001;8:1107–1121.
61. Matthews DA, Dragovich PS, Webber SE, et al. Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. Proc Natl Acad Sci USA 1999;96:11000–11007.
62. Barrette-Ng IH, Ng KK, Mark BL, et al. Structure of arterivirus nsp4. The smallest chymotrypsin-like proteinase with an alpha/beta C-terminal extension and alternate conformations of the oxyanion hole. J Biol Chem 2002;277:39960–39966.
63. Matthews BW. Solvent content of protein crystals. J Mol Biol 1968;33:491–497.
64. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004;60:2126–2132.
65. Ahmed FR, Rose DR, Evans SV, Pippy ME, To R. Refinement of recombinant oncomodulin at 1.30 Å resolution. J Mol Biol 1993;230: 1216–1224.
66. Kumar VD, Lee L, Edwards BF. Refined crystal structure of calcium-liganded carp parvalbumin 4.25 at 1.5-Å resolution. Biochemistry 1990;29:1404–1412.
67. Nakasako M. Large-scale networks of hydration water molecules around bovine beta-trypsin revealed by cryogenic X-ray crystal structure analysis. J Mol Biol 1999;289:547–564.
68. Nakasako M, Fujisawa T, Adachi S, Kudo T, Higuchi S. Large-scale domain movements and hydration structure changes in the active-site cleft of unligated glutamate dehydrogenase from *Thermococcus profundus* studied by cryogenic X-ray crystal structure analysis and small-angle X-ray scattering. Biochemistry 2001;40: 3069–3079.
69. Higo J, Nakasako M. Hydration structure of human lysozyme investigated by molecular dynamics simulation and cryogenic X-ray crystal structure analyses: on the correlation between crystal water sites, solvent density, and solvent dipole. J Comput Chem 2002;23:1323–1336.
70. Joti Y, Nakasako M, Kidera A, Go N. Nonlinear temperature dependence of the crystal structure of lysozyme: correlation between coordinate shifts and thermal factors. Acta Crystallogr D Biol Crystallogr 2002;58:1421–1432.
71. Nakasako M, Tsuchiya F, Arata Y. Roles of hydration water molecules in molecular packing of the killer toxin from *Pichia farinosa* in its crystalline state investigated by cryogenic X-ray crystallography. Biophys Chem 2002;95:211–225.
72. Mustata G, Briggs JM. Cluster analysis of water molecules in alanine racemase and their putative structural role. Protein Eng Des Sel 2004;17:223–234.
73. Henchman RH, McCammon JA. Extracting hydration sites around proteins from explicit water simulations. J Comput Chem 2002;23: 861–869.
74. Lounnas V, Pettitt BM, Phillips GN Jr. A global model of the protein-solvent interface. Biophys J 1994;66:601–614.
75. Badger J. Modeling and refinement of water molecules and disordered solvent. Methods Enzymol 1997;277:344–352.
76. Hope H. Introduction to cryocrystallography. In: Rossmann MG, Arnold E, editors. International tables for crystallography. Dordrecht: Kluwer Academic Publishers; 2001. p. 197–201.
77. Beamer LJ, Li X, Bottoms CA, Hannink M. Conserved solvent and side-chain interactions in the 1.35 Angstrom structure of the Kelch domain of Keap1. Acta Crystallogr D Biol Crystallogr 2005;61:1335–1342.
78. Xie P, Parsons SH, Speckhard DC, Bosron WF, Hurley TD. X-ray structure of human class IV sigmasigma alcohol dehydrogenase. Structural basis for substrate specificity. J Biol Chem 1997;272: 18558–18563.
79. Xie PT, Hurley TD. Methionine-141 directly influences the binding of 4-methylpyrazole in human sigma sigma alcohol dehydrogenase. Protein Sci 1999;8:2639–2644.
80. Guo J-T, Ellrott K, Chung WJ, Xu D, Passovets S, Xu Y.

PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction. Nucleic Acids Res 2004;32:W522–525.

81. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
82. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
83. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 2004;60:2256–2268.
84. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
85. Kleywegt GJ. Validation of protein models from Calpha coordinates alone. J Mol Biol 1997;273:371–376.
86. Brünger AT, Adams PD, Clore GM, et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54(Pt 5):905–921.
87. Jones TA, Zou JY, Cowan SW, Kjeldgaard. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 1991;47(Pt 2):110–119.
88. Martz E. Protein Explorer: easy yet powerful macromolecular visualization. Trends Biochem Sci 2002;27:107–109.
89. DeLano WL. The PyMOL Molecular Graphics System. San Carlos, CA: DeLano Scientific; 2002.
90. Axelrod HL, Feher G, Allen JP, et al. Crystallization and X-ray structure determination of cytochrome c_2 from *Rhodobacter sphaeroides* in three crystal forms. Acta Crystallogr D Biol Crystallogr 1994;50:596–602.
91. Young AC, Scapin G, Kromminga A, Patel SB, Veerkamp JH, Sacchettini JC. Structural studies on human muscle fatty acid binding protein at 1.4 Å resolution: binding interactions with three C18 fatty acids. Structure 1994;2:523–534.
92. Dunn CR, Banfield MJ, Barker JJ, et al. The structure of lactate dehydrogenase from *Plasmodium falciparum* reveals a new target for anti-malarial design. Nat Struct Biol 1996;3:912–915.
93. Liu Q, Huang Q, Teng M, et al. The crystal structure of a novel, inactive, lysine 49 PLA2 from *Agkistrodon acutus* venom: an ultrahigh resolution, ab initio structure determination. J Biol Chem 2003;278:41400–41408.
94. Rypniewski WR, Ostergaard PR, Norregaard-Madsen M, Dauter M, Wilson KS. *Fusarium oxysporum* trypsin at atomic resolution at 100 and 283 K: a study of ligand binding. Acta Crystallogr D Biol Crystallogr 2001;57:8–19.
95. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17:282–283.
96. Collaborative Computational Project. The CCP4 Suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr 1994;D50:760–763.
97. Schultze P, Feigon J. Chirality errors in nucleic acid structures. Nature 1997;387:668.