J. Mol. Biol. (1997) **265**, 445–464

JMB

Predicting Conserved Water-mediated and Polar Ligand Interactions in Proteins Using a K-nearest-neighbors Genetic Algorithm

Michael L. Raymer[1,2], Paul C. Sanschagrin[1], William F. Punch[2] Sridhar Venkataraman[1], Erik D. Goodman[2,3] and Leslie A. Kuhn[1]*

[1]*Protein Structural Analysis and Design Laboratory Department of Biochemistry* [2]*Genetic Algorithms Research and Applications Group Department of Computer Science and* [3]*Case Center for Computer-Aided Engineering and Manufacturing Michigan State University East Lansing, MI 48824 USA*

**Corresponding author*

Water-mediated ligand interactions are essential to biological processes, from product displacement in thymidylate synthase to DNA recognition by Trp repressor, yet the structural chemistry influencing whether bound water is displaced or participates in ligand binding is not well characterized. *Consolv*, employing a hybrid k-nearest-neighbors classifier/genetic algorithm, predicts bound water molecules conserved between free and ligand-bound protein structures by examining the environment of each water molecule in the free structure. Four environmental features are used: the water molecule's crystallographic temperature factor, the number of hydrogen bonds between the water molecule and protein, and the density and hydrophilicity of neighboring protein atoms. After training on 13 non-homologous proteins, *Consolv* predicted the conservation of active-site water molecules upon ligand binding with 75% accuracy (Matthews coefficient $C_m = 0.41$) for seven new proteins. Mispredictions typically involved water molecules predicted to be conserved that were displaced by a polar ligand atom, indicating that *Consolv* correctly assesses polar binding sites; 90% accuracy ($C_m = 0.78$) was achieved for predicting conserved active-site water or polar ligand atom binding. *Consolv* thus provides an accurate means for optimizing ligand design by identifying sites favored to be occupied by either a mediating water molecule or a polar ligand atom, as well as water molecules likely to be displaced by the ligand. Accuracy for predicting first-shell water conservation between independently determined structures was 61% (C_m=0.23). The ability to predict water-mediated and polar interactions from the free protein structure indicates the surprising extent to which the conservation or displacement of active-site bound water is independent of the ligand, and shows that the protein micro-environment of each water molecule is the dominant influence.

Keywords: hydration; drug design; solvent modeling; protein recognition; water site prediction

Introduction

Functional roles of bound water

It is increasingly recognized that water plays an important role in protein structure and function, yet the prediction of conserved protein-water interactions has remained elusive. In addition to bulk solvent, which is crucial to the hydrophobic effect and protein folding (Edsall & McKenzie, 1983; Tanford, 1980; Kuntz & Kauzmann, 1974), specific protein-bound water molecules have been shown to be important for substrate recognition in numerous protein structures. For example, examination of the crystal structure of the Trp repressor complex illustrates that the base-specific binding of operator DNA is achieved by a number of water-mediated hydrogen bonds between DNA bases and the protein (Joachimiak *et al*., 1994; Otwinowski *et al*.,

Abbreviations used: MHC I, class I major histocompatibility complex; PDB, Brookhaven Protein Data Bank; knn, k-nearest neighbors classifier; GA, genetic algorithm; adn, atomic density; ahp, atomic hydrophilicity; bval, temperature factor; hbd, number of hydrogen bonds; RMSD, root-mean-square positional deviation; bbknn, branch and bound k-nearest-neighbors classifier; PEG, polyethylene glycol.

0022–2836/97/040445–20 $25.00/0/mb960746

1988). In contrast, crystallographic structures for the class I human major histocompatibility complex (MHC I) with several peptidyl ligands (Wilson & Fremont, 1993) reveal that water can also allow plasticity in molecular recognition. In these structures, MHC I binds peptides of widely different side-chain chemistry using water-mediated contacts to bridge gaps between the protein and ligand. For thymidylate synthase, the crystal structure shows that a water molecule bound to absolutely conserved Tyr146 allows the enzyme to discriminate between substrate and product nucleotides (Fauman *et al*., 1994). The ubiquity of water-mediated ligand binding is reflected in a study of 20 non-homologous protein complexes, primarily with small ligands: the average protein–ligand interface includes 10 water molecules and 17 water-mediated bridges between protein and ligand (A. Cayemberg & L. A. Kuhn, unpublished results).

Bound water molecules also contribute to structural stability by forming extensive hydrogen-bond networks (Baker & Hubbard, 1984) and by lining grooves on solvent-exposed protein surfaces (Kuhn *et al*., 1992). Structures of proteins solved from crystals repeatedly rinsed in anhydrous organic solvent maintain the majority of their bound water molecules (Fitzpatrick *et al*., 1993; Travis, 1993), indicating that water is an integral part of the protein surface. Consideration of the structural and functional roles of bound water molecules has contributed to better drug design. For example, based on the knowledge that a specific water molecule (Wat301) is conserved in several X-ray structures of HIV-1 protease (Wlodawer *et al*., 1989), a higher-affinity cyclic urea inhibitor was obtained by incorporating a carbonyl oxygen to displace Wat301 and form the same network of hydrogen bonds (Lam *et al*., 1994).

When a ligand binds to a protein, each water molecule in the binding site will either be displaced by the ligand or remain bound, in both cases influencing the shape and energetics of interaction. Water molecules conserved in the ligand-bound structure generally participate in water-mediated hydrogen bonds between the protein and the ligand. The goal of *Consolv* is to examine the ligand-free protein structure and predict the conserved or displaced status of its water molecules upon ligand binding. For this purpose, bound water is defined as those crystallographically resolved water molecules in direct contact with the protein surface, using the criterion that the center of the water oxygen atom is within 3.6 Å of a protein atom center. *Consolv's* training and validation is for proteins without significant conformational change upon ligand binding. This is primarily because there are not enough known ligand-bound and free structural pairs for validation on proteins with conformational change upon ligand binding, and secondarily because it is difficult to define conserved sites between different conformational states. While ligand hydration (bound water) is likely to influence or contribute to water-mediated interactions, information on ligand-bound water is not included in *Consolv* for two reasons. First, a major application for the prediction of conserved or displaced active-site water is the design and optimization of protein inhibitors as drug candidates, and for this application the structures of the ligand and the protein–ligand complex are either unknown or subject to change. Secondly, for most protein complexes, the three-dimensional structure is not available for the free ligand, so the sites of ligand hydration are unknown. Therefore, neither the structure of the ligand nor of the protein–ligand complex is required for *Consolv's* decision procedure, and *Consolv's* accuracy indicates that active-site water conservation is reasonably ligand-independent.

Conservation of bound water under different crystallization conditions

Several studies have shown that many bound water sites are conserved for structures of a protein solved in different space groups, at different pH, in different salt concentrations, and even in different solvents. For instance, three structures of thermitase in complex with eglin-c and one without ligand were solved in the same space group from a range of crystallization conditions: pH 5.2 to 6.0, precipitants 5% to 25% PEG (polyethylene glycol) or 25% saturated ammonium sulfate, buffers 58 to 100 mM sodium acetate, 50 mM Bis-Tris, or 50 mM morpholinethane sulfonic acid, and calcium concentrations ranging from 0 to 100 mM. Superposition of these structures (Gros *et al*., 1992) shows that 18 water molecules are absolutely conserved in all four of the structures, and 38 are conserved in three of the four structures; for the three complexes with eglin-c, eight of the eleven active-site waters are always replaced by eglin-c, and three are always conserved. Structures of T4 lysozyme in seven space groups have been solved by one research group, with pH ranging from 6.1 to 8.5, salt concentrations of 0.1 to 2.2 M phosphate or ~0.25 M sodium acetate, alcohol concentrations from 0 to 5%, and 0 to 30% PEG. For the resulting 18 crystallographically independent lysozyme molecules, each with 38 to 141 bound water molecules, 50 to 60% of water sites are conserved; 23 water sites are found in at least nine of the eighteen structures (Zhang & Matthews, 1994). The 20 most frequently observed water sites are found in 62% of the structures, on average, replaced by another polar atom in an additional 3% of the structures, and displaced by a crystal contact in 14% of the structures. Zhang & Matthews consider this to be an underestimate of water conservation, because steric interference in the crystal lattice displaces some water sites which otherwise would be conserved, and because crystallographers tend to assign only water sites having high occupancy. A study of consensus

hydration sites in six complexes of FKBP12 with drug molecules shows that 60 of the 134 water binding sites are at least 50% conserved in the eight crystallographically independent structures, and 32 sites are conserved in at least three-quarters of the structures (Faerman & Karplus, 1995). In perhaps the most extreme change in crystallization conditions, the structure of subtilisin Carlsberg has been solved in an organic solvent, anhydrous acetonitrile, and compared with the structure solved in an aqueous environment (Fitzpatrick *et al*., 1993). Of the 119 bound water sites observed in the aqueous structure, 99 are conserved in the acetonitrile structure. Of the 12 subtilisin-bound acetonitrile molecules, four displace water molecules and four bind in the active site where eglin-c binds. Together, these studies show that one-half or more of bound water sites are typically conserved in independent structures of a protein solved under diverse conditions.

Solvent modeling and prediction

Theoretical approaches to solvent modeling have provided foundations for water refinement in crystallography (Jiang & Brünger, 1994; Badger, 1993) and for energy minimization approaches to solvent site prediction (Goodfellow & Vovelle, 1989; Wade *et al*., 1993; Zhang & Hermans, 1996). These methods employ potential functions and analysis of electron density to determine likely water binding sites for a protein structure. In contrast, empirical methods employ a database of information about known protein structures to make comparative predictions about water molecules in new structures. Examples are a neural network water site predictor based on residue chemistry and secondary structure (Wade *et al*., 1992); a method based on hydrogen-bond stereochemistry (Roe & Teeter, 1993); the *Auto-Sol* program based on hydrogen-bond directionality (Vedani & Huhta, 1991); and the *Aquarius2* method based on experimentally observed electron density and the distribution of water around protein residues (Pitt *et al*., 1993). The best methods achieve a predictive accuracy of 63 to 66% at protein surfaces, but are either restricted in applicability to polar residues only, or tend to overpredict hydration. Empirical methods, including the *Consolv* application presented here, are dependent on the quality of data included in the knowledge base, and thus can be subject to limitations in water fitting and refinement (Karplus & Faerman, 1994). Assignment of consensus water sites from superposition of independently solved X-ray structures (Faerman & Karplus, 1995) is a means of minimizing crystallographic artefacts in water structure. *Consolv* approaches this problem by training on water sites from a number of independently solved, non-homologous protein structures, and is the first empirical method developed to predict conservation of active-site water molecules upon ligand binding.

Algorithms

The *Consolv* knowledge base

The first step in the development of *Consolv* was the selection of protein structures to serve as a knowledge base for the decision process. The Brookhaven Protein Data Bank (PDB; Abola *et al*., 1987; Bernstein *et al*., 1977) was screened for proteins with independently solved ligand-bound and free structures; structures with a resolution $\leqslant$2.0 Å were preferred. To avoid statistical bias from inclusion of redundant information, molecular graphics screening and *Hera* hydrogen-bond diagrams (Hutchinson & Thornton, 1990) were used to cull structurally related proteins from the knowledge base. To exclude structures with conformational and chemical differences between the ligand-bound and free structures that could affect conservation of bound water (hydration), we included only those pairs with no sequence variations, mutations, or significant backbone conformational changes near the active site, and low main-chain root-mean-square positional deviation (RMSD $\leqslant$ 1.0 Å) upon superposition of the ligand-bound and free structures using *InsightII*

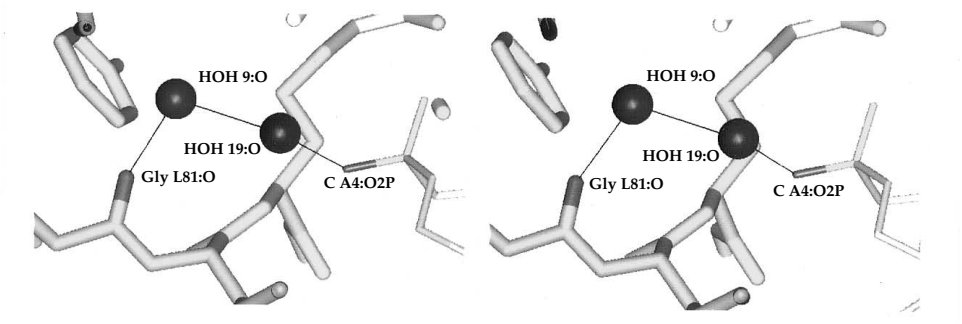


Figure 1. Stereo view of a di-water bridge linking barnase (thick tubes at left) with its tetranucleotide ligand (thin tubes at right) in PDB structure 1brn. The bridging water molecules are indicated by dark spheres, and hydrogen bonds (each $\leqslant$3.6 Å) spanning the bridge are shown by thin lines. This figure was rendered using *InsightII* (Molecular Simulations, Inc., San Diego, CA).

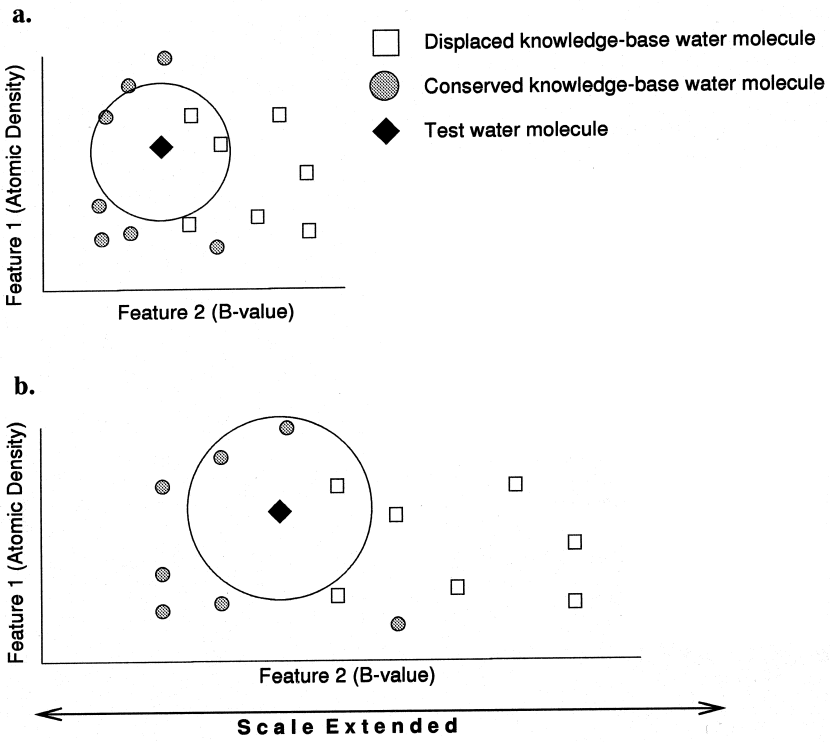


Figure 2. Scaling of knn feature axes. Increasing the scale of the *x*-axis (*B*-value) takes advantage of the relatively greater discrimination ability of *B*-value relative to atomic density (*y*-axis, not rescaled) in distinguishing conserved from displaced water molecules. Scaling the *B*-value axis between (a) and (b), using a weight selected by the genetic algorithm, can change the predicted status for a test water (shown here to change from displaced to conserved status). Upon axis scaling, the radius of the circle representing the neighborhood in the knn algorithm changes as needed to encompass exactly *k* neighbors.

software (Molecular Simulations, Inc., San Diego, CA). For all analyses, the active site was defined as the intersection of a 3.6 Å envelope around the protein and a similar envelope about the ligand; the active site of the free protein structure was identified through structural superposition with the ligand-bound form. The initial protein set, compiled according to the above criteria, consisted of 13 structural pairs (Table 1), representing a number of distinct structures with a variety of biological functions. These proteins bind diverse ligands, including lipids, small organic molecules, peptides, and DNA oligomers.

The primary knowledge base for *Consolv*, consisting of first hydration shell and active-site water molecules and measurements of their protein environments, was compiled from this initial set of 13 structural pairs. Water molecules within 3.6 Å of protein surface atoms, thus capable of making van der Waals' contacts or hydrogen bonds to atoms in the protein, were considered to be first-shell waters. (This and other distance criteria are from atom center to atom center.) The 3.6 Å threshold also includes the major peak in the radial distribution function of protein-associated water (Kuhn *et al.*, 1995). Active-site water molecules, identified above, included all water molecules potentially participating in protein-ligand interactions. Water bridges between protein and ligand sometimes involved a single water molecule, and in other cases involved two water molecules, comprising a "di-water bridge" (Figure 1) of the form: (protein atom)–(water molecule)–(water molecule)–(ligand atom), where each indicated interaction is a hydrogen bond of $\leqslant$3.6 Å. Therefore, each water molecule within 3.6 Å of both the protein and the ligand was considered an active-site water, and waters possibly participating in di-water bridges were identified by including all other water molecules within 3.6 Å of a first-shell active-site water. To maintain computational tractability, 1700 first-shell water molecules, including both active-site and non-active-site waters, were selected from the 13 ligand-free structures and included in the knowledge base. This knowledge base included all

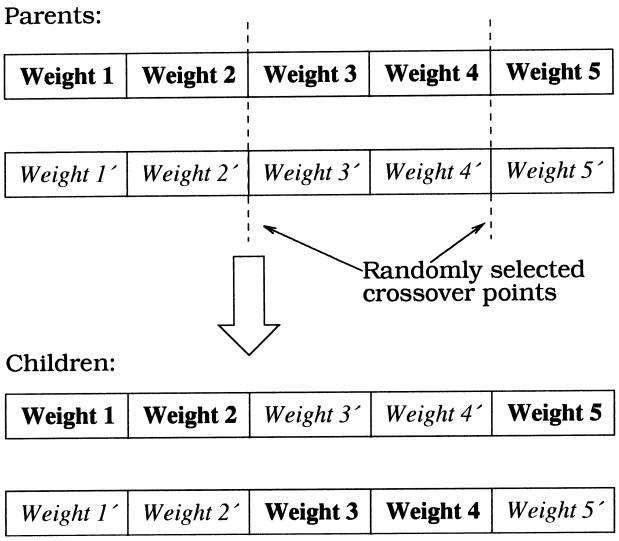


Figure 3. The two-point crossover operation combines two parent weight sets to form two new sets by interchanging weights occurring between the crossover points.

850 waters determined to be displaced in the corresponding ligand-bound structures, and a randomly-selected set of 850 waters determined to be conserved. The overlapping set of 157 active-site water molecules was used for training *Consolv* to optimize prediction on active-site waters.

The next step in development of the knowledge base was determination of the conserved or displaced status of each of its water molecules. For each protein, the main chain of the ligand-bound structure was superimposed onto that of the free structure using *InsightII* software. A computer program was written to identify equivalent water sites in the free and ligand-bound protein structures by using each water molecule in the free structure as a reference site, superimposing the ligand-bound structure, and identifying any water molecules in the ligand-bound structure with their oxygen atoms within 1.2 Å (Zhang & Matthews, 1994) of the reference water oxygen from the free structure. Given that the effective radius of a water molecule including hydrogen atoms is 1.4 to 1.6 Å, if two water sites have oxygen atoms 1.4 to 1.6 Å apart, they will contact each other. Thus, a 1.2 Å distance between oxygen atoms in two waters results in considerable overlap and provides a conservative criterion for defining equivalent water sites in the superimposed structures. It would be unexpected to find two water sites in the complex that superimpose to within 1.2 Å of a water in the free protein (this would imply that the two waters from the complex were at most 2.4 Å apart, too close for hydrogen bonding; Faerman & Karplus, 1995); however, in such cases, the closer of the two was considered to correspond to the water site in the free structure, while the more distant water molecule was left for possible assignment to another water from the free structure. Using this approach, the water molecules in the knowledge base (each from a ligand-free structure) were identified as conserved or displaced in the corresponding ligand-bound structure.

Measurement of bound water environment

Four features were selected to represent the microenvironment of each water molecule in the knowledge base and serve as a basis for prediction of conserved water sites. The first feature was atomic density, defined as the number of protein atoms within 3.6 Å of the water molecule, providing a measure of protein surface topography. Deep grooves in the protein surface have higher atomic density values than convex protein surfaces, and water binding is twice as frequent in protein grooves as on flat or protruding surface regions (Kuhn *et al*., 1992). The next feature, atomic hydrophilicity, measures the tendency of surrounding atoms to bind water molecules, based on a study of the frequency of hydration for each atom type in 56 high-resolution protein structures (Kuhn *et al*., 1995). For each water molecule, the atomic hydrophilicity values of all protein atoms and water molecules within 3.6 Å were summed and stored in the knowledge base. The third environmental feature was the number of hydrogen bonds between the water molecule and protein atoms, evaluated using the program *Hbond* (Overington *et al*., 1990), which identifies hydrogen bonds based on the occurrence of donor and acceptor atoms within 3.5 Å. The hydrogen-bonding capacity of protein atoms correlates highly with the frequency of hydration (Baker & Hubbard, 1984; Kuhn *et al*., 1995). The fourth environmental feature was the crystallographic temperature factor (*B*-value, measured in $Å^2$), as reported in the PDB entry for the protein, which provides a measure of thermal mobility of the water molecule, as well as local disorder in the crystal lattice. While *B*-values may be imprecise and refinement dependent, they provide a relative indication of atoms' thermal mobility that reflects the tendency of a water site to be conserved between structures (Karplus & Faerman, 1994).

Taken together, these features provided a characterization of the micro-environment of each water molecule incorporating three features known to correlate with water binding: atomic density, number of hydrogen bonds, and temperature factor. The fourth feature, atomic hydrophilicity, is highly correlated with atomic density and hydrogen bonding and was included because it might provide predictive ability equivalent to these two other features, allowing a reduction in the number of features used for classification. Other features were not assessed; some which may also be useful for water site evaluation are described in the section on *Consolv* enhancements. The feature characterization described above serves as the basis for comparing the environments of known conserved or displaced water molecules with the environments of test water sites being evaluated for their likelihood of conservation.

The *Consolv* method

Consolv's decision procedure uses a previously developed algorithm coupling a *k*-nearest-neighbors classifier (knn) with a genetic algorithm (Punch *et al*., 1993). To apply the knn classifier to predict conserved water sites, the four features in the knowledge base (atomic density, atomic hydrophilicity, number of hydrogen bonds, and temperature factor) were used as the axes in four-dimensional space. Each knowledge-base water had known conserved or displaced status, and was plotted in this four-dimensional space based on its value for each feature. Finally, a test water molecule of unknown status was plotted among these knowledge base waters based on its four feature values, and in our implementation, the *k* nearest neighbors (for some small, positive integer *k*) from the knowledge base voted to predict the category of the test water. If a majority of the voting waters belonged to the conserved water category, the knn predicted that the unknown water was also

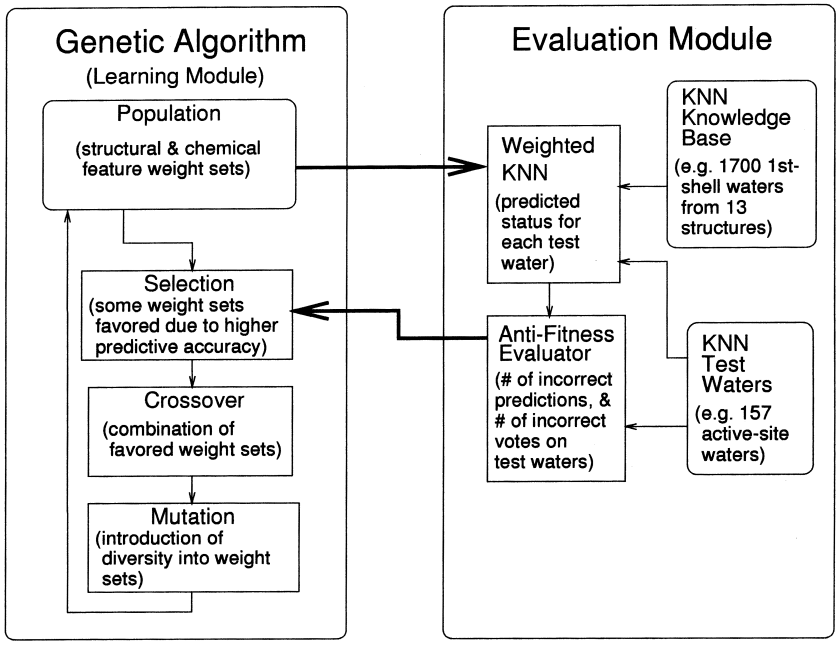


Figure 4. Training of *Consolv* for active-site water prediction. Feature weight sets are passed to the evaluation module, where they are used to classify the test waters (for which the conserved or displaced status is known). Anti-fitness, based on the number of incorrect votes and incorrect predictions on these test waters, is returned to the GA to aid in selection of weight sets for the next knn evaluation. 100 or more such cycles (generations) are executed in a training session, iteratively improving the weight set for optimally categorizing the test waters. The population typically consists of 200 to 500 weight sets tested by the knn in each generation.

a conserved water molecule. Odd values of k are typically used when there are an even number of categories (as in our case, conserved/displaced), to avoid the possibility of a tie.

Knn techniques are commonly employed for analyzing data sets which cannot be assumed to follow a normal distribution, and have been applied to protein secondary structure prediction (Yi & Lander, 1993; Salamov & Solovyev, 1995). Despite their utility, pure knn classifiers are susceptible to noisy input data, outliers, and spurious or correlated selection features. Fortunately, the knn can be tuned to overcome these limitations. One method for increasing its power and robustness is by weighting the feature axes according to their relative importance (Kelly & Davis, 1991; Siedlecki & Sklansky, 1989; Punch *et al*., 1993). For example, if it is known in advance that one feature (e.g. temperature factor) is more relevant to water site conservation than the others, the scale of the axis associated with this feature can be increased to heighten the ability of the knn algorithm to discriminate between water categories along this axis (Figure 2). Utilization of such a weighted knn presupposes that there is some *a priori* knowledge of the relative contribution of each feature. In the case of conserved water site prediction, no such information is available. In fact, a deeper understanding of the factors which contribute to conservation of water sites between structures is an important objective. *Consolv*, which combines this weighted knn with a genetic algorithm for testing different axis weights, was used to optimize the prediction of conserved water sites. The genetic algorithm's task was to find the set of axis weights that allowed the weighted knn algorithm to achieve improved prediction rates using a given knowledge base. These weights could then be applied using the fast knn algorithm, without the genetic algorithm, for predicting the status of water molecules.

Genetic algorithms (GAs) are learning procedures modeled after the mechanics of Darwinian natural selection and evolution (Goldberg, 1989; Holland, 1975). Their ability to solve deceptive, multimodal, high-dimensionality problems has proven GAs to be effective problem solvers in many areas of biochemistry, including protein conformation and folding (Patton *et al*., 1995; Dandekar & Argos, 1994; Unger & Moult, 1993; Le Grand & Merz Jr, 1994; Ring & Cohen, 1994), structural alignment and comparison (May & Johnson, 1994), evolution of receptor models (Walters & Hinds, 1994), and ligand docking (Jones *et al*., 1995; Oshiro *et al*., 1995). A distinguishing feature of a GA is maintenance of a large population of potential solutions, each in competition with the others. The initial population is simply a randomly generated sample of possible solutions. These potential solutions are referred to as individuals, and represented as a string of characters, or chromosome. Each generation, the individual solutions are rated by a function measuring their fitness, or how well they solve the problem (in this case, prediction of conserved water sites). Highly fit individuals are more likely to be selected for inclusion in the next generation, while individuals with low fitness have a small, but non-zero, chance of being selected. The selected individuals may then be modified by one or more genetic operators before advancing to the next generation. Typically, after a number of generations, the individuals in the population will begin to converge towards a single, near-optimal solution.

The genetic operators employed are crossover and mutation. As in biology, crossover combines

the ''genetic material'' of two parents to create offspring. In GAs, a crossover consists of a simple substring swap between two chromosomes (Figure 3) and results in two recombinant offspring. In our application, each chromosome consisted of four contiguous real numbers (each represented by 32 bits), which were used as feature weights in the knn algorithm. Crossover allows the GA to learn by recombining parts of high-quality solutions (chromosomes) to produce possibly better solutions (Goldberg, 1989). A mutation involves making a random change to a chromosome. At its most basic level, information in a genetic algorithm (in our case, a set of feature weights) is represented as a string of binary digits, and the mutation operator selects a random binary digit of the chromosome and inverts it (0 to 1, or 1 to 0). The result of such a ''point mutation'' is a random change in one of the four feature weights. Since the crossover operator is not constrained to cut the chromosome along the weight boundaries, crossover can split a single axis weight among two offspring, effectively causing a mutation to each of the resulting chromosomes. After selection, crossover, and mutation at the beginning of each generation, the feature weights from each individual in the new population are used to scale the axes of the knn classifier.

Crossover is the primary learning mechanism of the GA, while mutation maintains population diversity and prevents convergence to local optima. Through manipulation of the rates of crossover and mutation, GA behavior can be balanced between rapid search and broad coverage of the search space. In our experiments, the probability of crossover was set to a value between 0.6 and 0.8 crossovers per individual per generation, and mutation rates between 0.01 and 0.0001 mutations per bit were tested.

Consolv's fitness function

In designing a function to compute the fitness of each weight set, the function should be as smooth as possible to facilitate effective search by the GA. If fitness were based only on the number of correctly classified test water sites, the function would increment in discrete steps for each correctly predicted water, and between these discrete steps there would be no guidance to the GA on how the weights should be modified to increase the number of correct predictions. However, increasing the number of correct knn votes for the status of each water molecule would eventually increase the number of correct predictions, so including a

Table 1. Non-homologous crystallographic structural pairs (ligand-bound and free) selected for the *Consolv* knowledge base

Protein	PDB Code	Source	First-shell water molecules	Active-site water molecules	Resolution (Å)	*R*-factor	RMSD[a] (Å)
Rhizopuspepsin	2APR	*Rhizopus chinensis*	181	20	1.8	0.143	
Complex w/peptide inhibitor	3APR		163	11	1.8	0.147	0.13
Chloramphenicol acetyltransferase	2CLA	*Escherichia coli*	88	7	2.35	0.152	
Complex w/chloramphenicol	3CLA	Strain rz1032	185	9	1.75	0.157	0.41
Proteinase a	2SGA	*Streptomyces griseus*	184	18	1.5	0.126	
Complex w/tetrapeptide	5SGA		156	4	1.8	0.116	0.08
Thermitase	1THM	*Thermoactinomyces*	185	18	1.37	0.166	
Complex w/eglin-c	2TEC	*vulgaris*	183	18	1.98	0.165	0.24
Thermolysin	3TLN	*Bacillus thermoproteolyticus*	153	8	1.6	0.213	
Complex w/Val-Trp	3TMN		157	9	1.7	0.173	0.10
Actinidin	2ACT	*Actinidia chinensis*	200	16	1.7	0.180	
Complex w/e-64	1AEC		196	11	1.86	0.145	0.11
Adipocyte lipid-binding protein	1LIB	*Mus musculus*	83	6	1.7	0.180	
Complex w/hexadecanesulfonic acid	1LIC		64	2	1.6	0.195	0.32
Barnase	1BSA	*Bacillus amyloliquefaciens*	240	17	2.0	0.173	
Complex w/DNA (CGAC)	1BRN		208	23	1.76	0.190	0.45
Bira bifunctional protein	1BIA	*Escherichia coli*	43	3	2.3	0.190	
Complex w/biotinylated lysine	1BIB		19	1	2.8	0.173	0.48
Carboxypeptidase A	5CPA	*Bos taurus*	194	14	1.54	0.190	
Complex w/phosphonate	6CPA	Pancreas	127	5	2.0	0.193	0.36
Deoxyribonuclease I	3DNI	*Bos taurus*	221	11	2.0	0.177	
Complex w/DNA	2DNJ	Pancreas	210	19	2.0	0.174	0.37
Cholesterol oxidase	3COX	*Brevibacterium*	409	13	1.8	0.156	
Complex w/dehydroisoandrosterone	1COY	*Sterolicum*	367	1	1.8	0.159	0.24
Carbonic anhydrase II	1CA2	*Homo sapiens*	152	4	2.0	0.173	
Complex w/trifluoromethane-sulphonamide	1BCD	Erythrocytes	193	6	1.9	0.154	0.20

[a] Root-mean-square positional deviation for all protein backbone atoms between the ligand-bound and free structures.

correct-vote term in the fitness function resulted in a smoother function and drove the GA to change the weights in a direction that improved predictions. Thus, the fitness score for each weight set had two parts: one measuring the number of correct predictions, and another measuring votes for the correct category.

The GA engine for *Consolv*, based on GAUCSD code (Schraudolph & Grefenstette, 1992), is designed to minimize, rather than maximize, a function. Therefore, *Consolv's* GA minimized anti-fitness, measured by incorrect votes and predictions:

$$\text{anti-fitness}(\textit{weight set}) = \begin{pmatrix}\%\text{ incorrect predictions}\\ \text{in each category}\end{pmatrix} + \begin{pmatrix}\%\text{ incorrect}\\ \text{votes}\end{pmatrix} = \left(\sum_{i=1}^{c} \frac{p_i}{t_i} \times \frac{1}{c}\right) + \left(\frac{v}{nk}\right)$$

where:

c = the number of categories = 2 (conserved or displaced),
p_i = the number of incorrect predictions for category i,
t_i = the number of waters of category i,
v = the total number of incorrect votes,
k = the number of voting neighbors (the k parameter for the knn), and
n = the total number of waters being evaluated.

For some of the training runs, a weight parsimony term was included in the fitness function to find the minimum magnitude for each weight consistent with high predictive accuracy. The parsimony term equalled the average of the four feature weights multiplied by a parsimony constant, and resulted in a linear penalty for increasing any of the weights. A parsimony constant of 0.04 reduced feature weights to values generally less than 2, maintained consistency between weights in independent GA runs, and kept predictive accuracy to within one or two percent of the accuracy in the absence of a parsimony term. By the end of a typical GA run, a parsimony term with a constant of 0.04 contributed 3% to the overall fitness function.

Consolv training and testing

Training of the knnGA for *Consolv*, shown schematically in Figure 4, consisted of a two-way interaction between the weighted knn prediction module and the GA learning module. The GA passed feature weights to the knn, and the weights were used to linearly scale the corresponding knn axes. This scaled knn feature space (Figure 2) was then used to make predictions for a test set of water molecules. Since the conserved or displaced status of these test water molecules was known in advance, the knn could return a fitness score for each weight set based on the number of correct predictions achieved using a particular set of axis weights. The GA used these fitness scores as the basis for selecting weight sets to proceed to the next generation. Upon convergence of the knnGA algorithm, this training produced an optimized weight set, which could then be used by a single run of the weighted knn algorithm without the GA to predict the conserved or displaced status of water molecules in new structures not included during training. The training process could either be halted after a fixed number of knnGA generations, or when the incremental gains in predictive accuracy dropped below a threshold. *Consolv* training runs were executed for a fixed number of generations ranging from 100 to 500; most runs proceeded for 200 generations. In all cases, the incremental prediction gains became negligible before the run terminated. Population sizes ranged from 200 to 500 individuals (weight sets), for a total of 20,000 to 250,000 knn executions per training run.

The GA represents feature weight sets as a linear string on a chromosome, and weights placed together on the chromosome are less likely to be split during crossover than weights farther apart. Thus, values which are nearby on the chromosome may be cooptimized, or ''linked''. For chromosomes with a large number of features, linkage can be minimized using an additional genetic operator called inversion. The inversion operator reverses the order of a randomly selected segment of the chromosome. The original configuration of the segment is tracked, so the change does not affect the phenotypic expression of the chromosome, merely the way in which it is stored in the GA. Using inversion, every pair of features on the chromosome is equally likely to be nearby, or linked, during the course of a GA run. For *Consolv*, the number of feature weights on the GA chromosome was not large enough to warrant including an inversion operator. To assess possible linkage, several GA runs were conducted using different chromosome orderings (''static inversion''), while keeping all other GA parameters the same.

Consolv was trained on the 1700 first-shell water molecules selected for the knowledge base. Training experiments were designed to optimize the performance of *Consolv* on the prediction of conserved or displaced status for active-site water molecules in the absence of ligand knowledge, or the prediction of conservation of first-shell water sites between independently-solved structures. For all runs, the environmental features in the database were normalized to range over the continuous closed interval [1.0 to 10.0] to eliminate implicit axis weighting resulting from different ranges of values for the four different features. After training was complete, *Consolv* was tested on the active-site waters from seven new, non-homologous proteins (Table 2) chosen according to the criteria used for the original 13 structures. *Consolv's* knn module was run on these new proteins using the feature

Table 2. Seven new crystallographic structural pairs selected for unbiased testing

Protein	PDB Code	Source	First shell waters	Active site waters	Resolution (Å)	*R*-factor	RMSD[a] (Å)
Glutathione *S*-transferase	[b]	*Schistosoma*	94	3	2.4	0.197	
Complex w/praziquantel		*japonica*	64	1	2.6	0.212	0.19
RTEM-1 β-lactamase	[c]	*Escherichia*	164	6	1.7	0.184	
Complex w/penicillin G		*coli*	176	4	1.7	0.182	0.22
Cyclodextrin glycosyltransferase	1CGT	*Bacillus*	532	9	2.0	0.187	
Complex w/glucose	1CGU	*circulans*	410	8	2.5	0.166	0.34
Enolase	3ENL	*Saccharomyces*	279	5	2.25	0.154	
Complex w/2-phospho-D-glyceric acid	5ENL	*cerevisiae*	304	14	2.2	0.148	0.21
Trp repressor	2WRP	*Escherichia*	142	32	1.65	0.180	
Complex w/synthetic operator	1TRO	*coli*	250	66	1.9	0.167	2.18
Concanavalin A	2CTV	*Canavalia*	140	7	1.95	0.153	
w/α-methyl-D-mannopyranoside	5CNA	*ensiformis*	158	7	2.0	0.199	0.42
Dihydrofolate reductase	1DR2	*Gallus gallus*	71	6	2.3	0.158	
Complex w/biopterin	1DR3	Liver	100	16	2.3	0.140	0.12

[a] Root-mean-square positional deviation for all protein backbone atoms between the ligand-bound and free structures.
[b] Provided by Drs Michele McTigue and John Tainer, The Scripps Research Institute (McTigue *et al*., 1995).
[c] Provided by Drs Natalie Strynadka and Michael James (Strynadka *et al*., 1992).

weight sets evolved from training on the 1700-water knowledge base (a balanced number of conserved and displaced sites from the 13 structures in Table 1) or the 2832-water knowledge base (a balanced number of conserved and displaced sites from all 20 structures in Tables 1 and 2). These tests were then evaluated to determine the method's accuracy.

Accuracy was measured in three ways: by the number of correctly predicted water sites out of the total number of evaluated sites (percentage accuracy), by the percentage accuracy for each class (conserved, displaced), and by the Matthews coefficient C_m, which assesses the balance of accuracy between classes (Matthews, 1975). Using the abbreviations P for the number of predicted sites, O for observed, C for conserved, and D for displaced (e.g. PCOC = ''number of sites predicted conserved and observed conserved''):

$$C_m = \{(PCOC)(PDOD) - (PDOC)(PCOD)\}$$
$$\div \{(OC)(PC)(OD)(PD)\}^{1/2}$$

While C_m is useful for assessing accuracy and balance simultaneously, it can have a small value even when the accuracy for each class is high. This occurs, for instance, when the number of observed members for the two classes is significantly different. Furthermore, C_m is undefined (equals infinity) if one of the classes has no observed members, or if it has no predicted members.

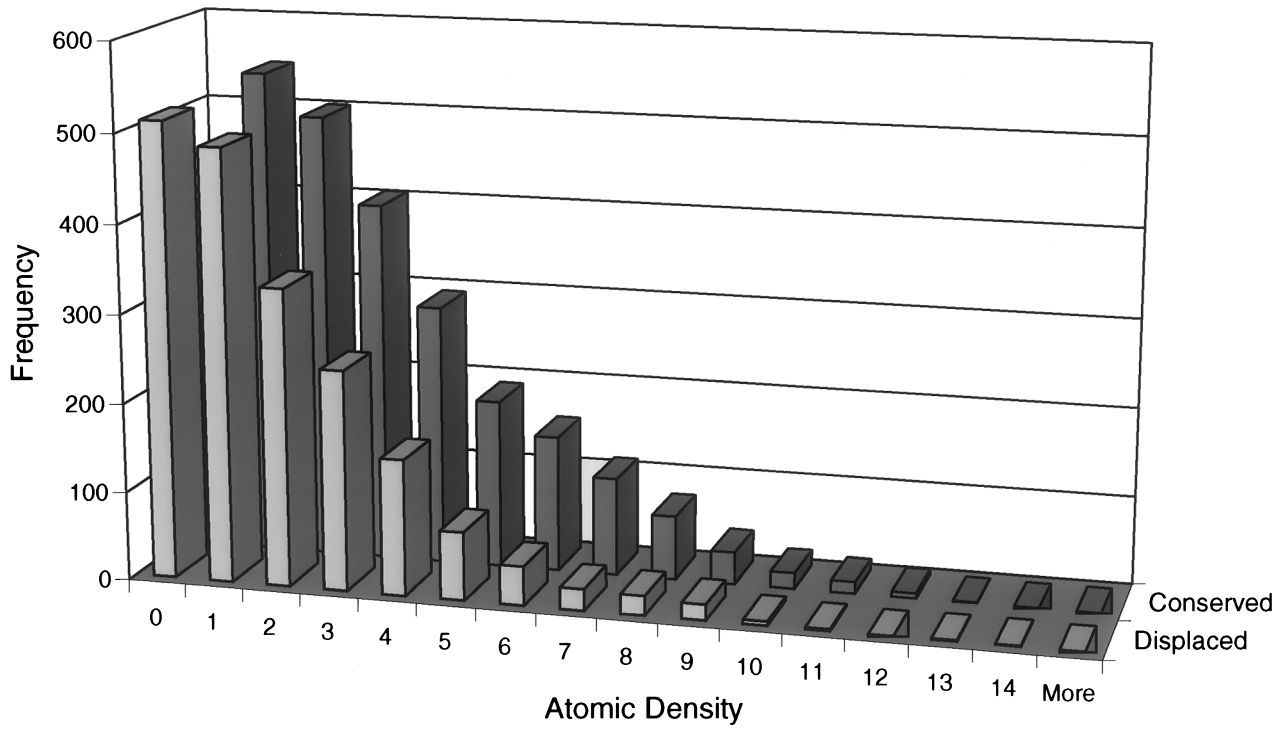
To test the ability of standard statistical methods to differentiate between environments associated with conserved and displaced water sites, the *Discrim* module of the *SAS* statistical analysis package (SAS Institute, Inc., Cary, NC) was used to perform discriminant analysis on the 1700-water data set plotted in the four-dimensional environmental feature space, and the results were compared with those of *Consolv*. Discriminant analysis identifies the axis (in general, a linear combination of the original axes) along which the classes of objects (e.g. conserved and displaced) can be maximally separated. In addition, threshold tests were performed to determine whether a decision boundary perpendicular to a feature axis (e.g. a decision rule such as ''if *B*-value >70.0 Å^2, then predict displaced'') could be constructed such that nearly all objects above the boundary belonged to the same class, providing good water classification. A program was written to extract all water molecules from the *Consolv* first-shell knowledge base meeting a given threshold criterion (e.g. all waters with *B*-value >70.0 Å^2). Thresholds were finely sampled for each parameter, and the proportion of conserved and displaced waters above each threshold was examined to determine if conserved or displaced waters predominated.

Results and Discussion

Statistics of conserved and displaced water molecules

Statistical analysis of first-shell water molecules in the 13 training proteins (Table 1) revealed some interesting trends. The ligand-free structures contained a total of 2334 first-shell water molecules. Of these, 850 were displaced in the corresponding ligand-bound structures, while 1484 were conserved, resulting in a displaced:conserved ratio of 1:1.75. *Consolv's* 1700-water knowledge base included the largest balanced set of 850 displaced and 850 conserved water sites from these proteins. The displaced:conserved ratio was reversed in the active sites, where ligand interactions can displace water. There were 157 active-site bound water molecules in the ligand-free structures, 114 displaced and 43 conserved, resulting in a displaced:conserved ratio of 2.7:1. Distribution

a.



b.

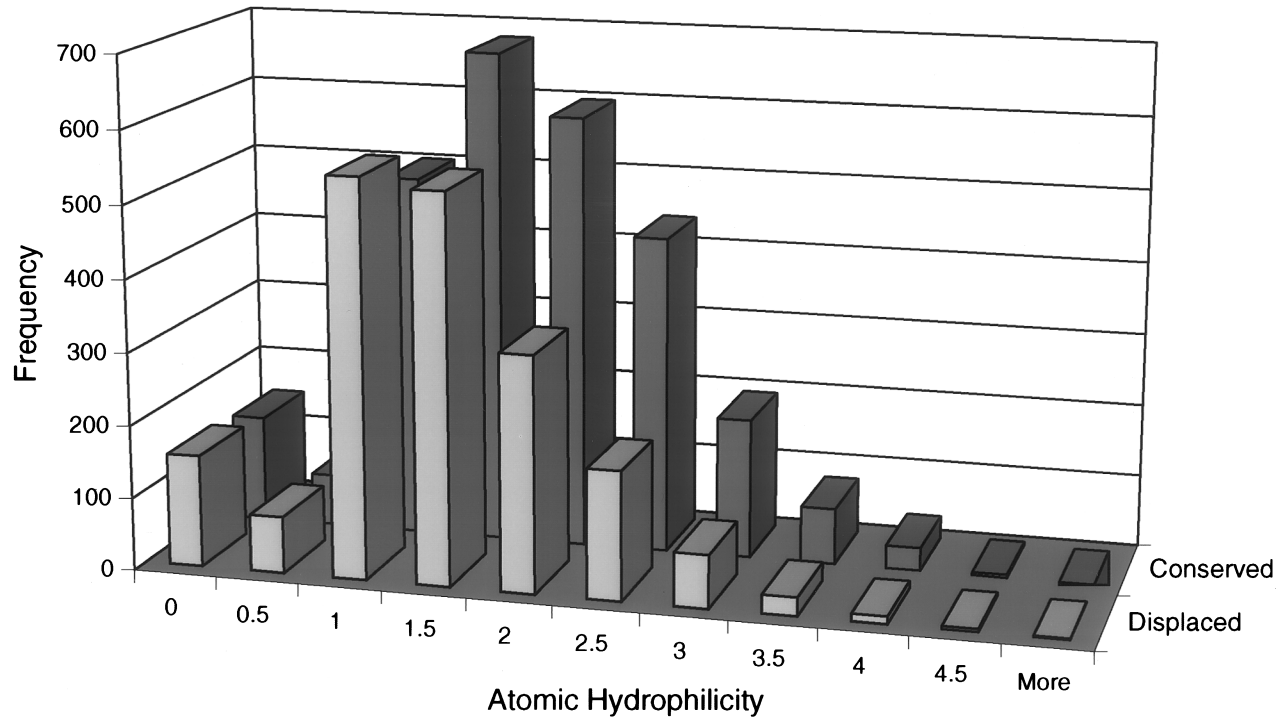
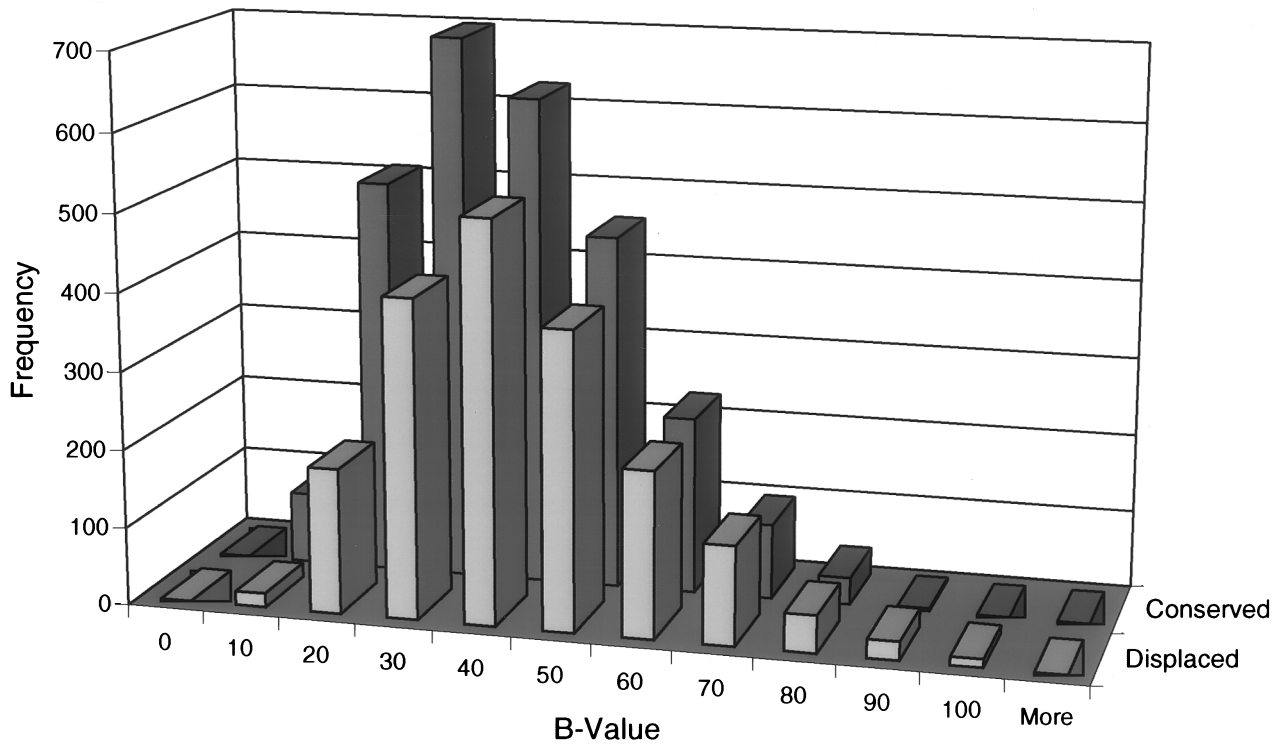


Figure 5(a and b) *legend opposite*

c.



d.

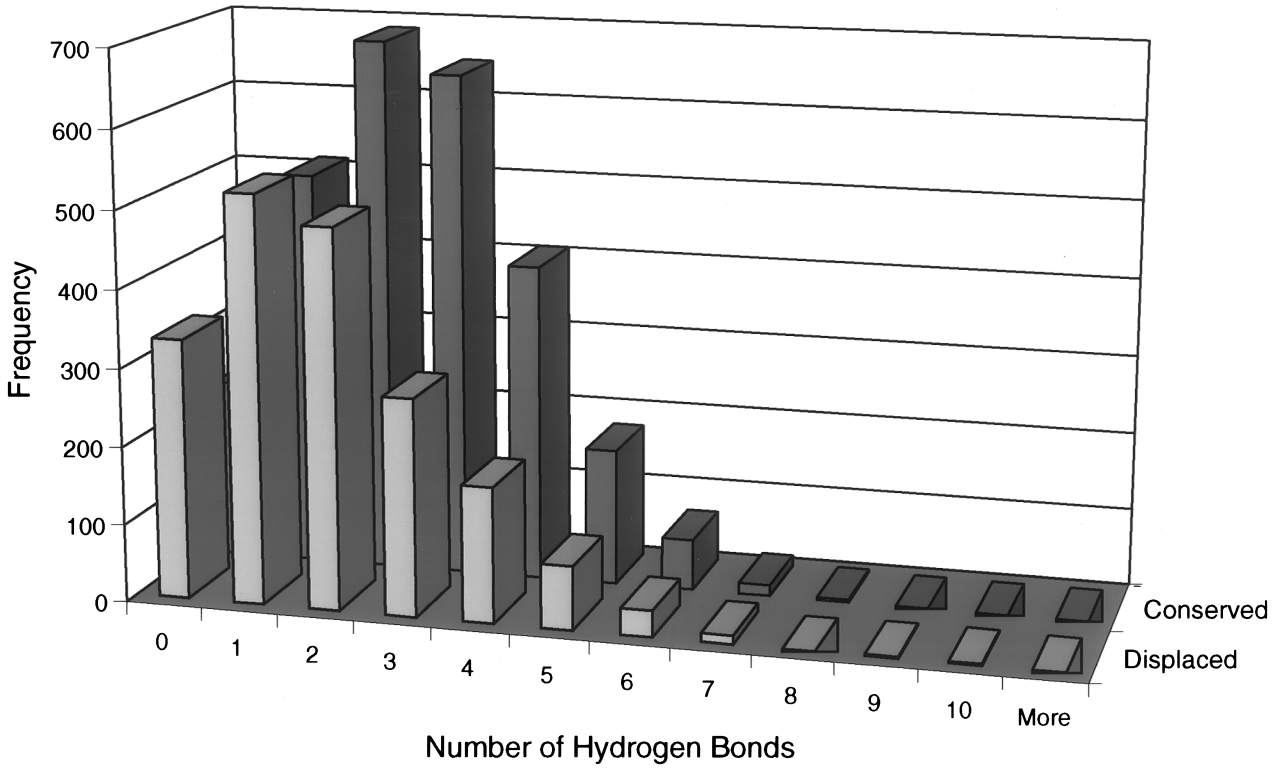


Figure 5. Distribution of environmental feature values for conserved and displaced water sites. The unnormalized values of atomic density, atomic hydrophilicity, temperature factor, and number of hydrogen bonds are shown separately for all conserved and all displaced (non-conserved) first-shell water sites in 20 pairs of free and ligand-bound protein structures (Tables 1 and 2). The histograms show the distribution of the features, measured in the ligand-free structure as follows: (a) atomic density, the number of protein atoms within 3.6 Å; (b) atomic hydrophilicity, the sum of the expected number of hydrations (Kuhn *et al*., 1995) for all protein atoms and water molecules within 3.6 Å; (c) *B*-value, the PDB temperature factor for the water molecule (Å^2); and (d) the number of hydrogen bonds between the water molecule and protein donor or acceptor atoms within 3.6 Å.

functions for atomic density, atomic hydrophilicity, temperature factor, and the number of hydrogen bonds for conserved and displaced water molecules are presented in Figure 5. The distributions for conserved and displaced sites for each feature were highly overlapping, presenting a challenge for classification and suggesting the importance of evaluating the features simultaneously to provide more information for distinguishing conserved from displaced sites.

Training results

Consolv's GA parameters were individually optimized through a number of preliminary experiments. One of the most sensitive search parameters proved to be the k value used by the knn module, determining the number of neighbors that vote on the status of each new water molecule. The effect of varying k was analyzed using an unweighted version of the knn algorithm (a computationally efficient alternative to the knnGA), allowing tests of a number of k's. The k-dependence of predictive accuracy was evaluated for two datasets: the 157 active-site waters, and the 1700 first-shell waters (Figure 6). Results of these tests showed that $k = 3$ is favored for active-site water molecules, whereas a larger k value ($k = 39$) is favored for first-shell water molecules. Similar tests using the *Consolv* knnGA for $k = 3$, 5, and 7 indicated that a k value of 3 consistently outperformed other k's tested for active-site prediction. Odd values of k were used to avoid invoking tie-breaking schemes. Additional work remains to determine the asymptote in predictive accuracy as a function of k for first-shell waters using a weighted knn. However, from this unweighted nearest-neighbor analysis, it appears that clusters of conserved or displaced active-site water molecules with similar structural and chemical environments contain only ~ 3 water molecules, whereas clusters of conserved or displaced first-shell waters with similar environments contain ~ 39 water molecules.

Initial testing also showed that balancing the number of conserved and displaced waters in the knowledge base improved the accuracy of predictions. In the first hydration shell of knowledge-base proteins, there were almost twice as many conserved water sites as displaced. Since the knn prediction is based on the categories of voting waters from the knowledge base, balancing the two categories was necessary to avoid a strong bias towards prediction of conserved waters. In a recent application of a knn algorithm toward secondary structure prediction, this balance was realised by weighting the categories in the voting procedure (Salamov & Solovyev, 1995). We achieved similar results by including equal numbers of conserved and displaced water molecules in the knowledge base. Before balancing the knowledge base, *Consolv's* predictive accuracy for displaced waters from the 157 active-site waters test set was 95%, and

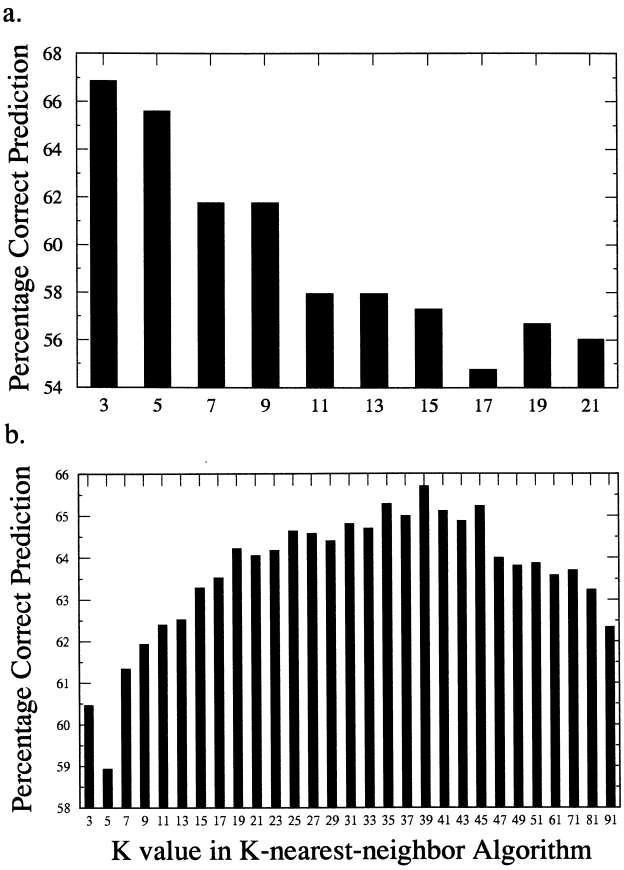


Figure 6. An unweighted knn algorithm was applied to waters in the knowledge base, using a range of k values to determine the optimal k for water site classification. 157 active-site or 1700 first-shell water molecules were classified by *Consolv* as conserved or displaced between independently solved structures, based on normalized values for atomic density, hydrophilicity, temperature factor, and the number of hydrogen bonds at each water site. When classifying active-site water molecules (a), small values of k (~ 3) yielded better predictions. In contrast, higher k values (~ 39) were more effective for classifying first-shell waters (b).

for conserved waters, 22%. After balancing the knowledge base to contain an equal number of conserved and displaced water molecules (850 of each), accuracy for conserved waters was 83%, while accuracy for displaced waters was 75%.

In these preliminary training runs, used to optimize feature weights for later unbiased tests, the weight sets were applied to predict the status of 1700 first-shell water molecules in the knowledge base or 157 active-site water molecules from the proteins in Table 1. All 114 of the displaced waters and a random sampling of conserved waters from the 157 active-site waters were included in the knowledge base, due to its being constructed to contain a maximal, equal number of conserved and displaced first-shell water sites. This overlap between the knowledge base and the waters being tested facilitated a self-consistency check for *Consolv*, as well as providing a set of feature weights for use in later unbiased testing. The

Table 3. Results of *Consolv* knnGA training and knn testing on active-site and first-shell water molecules

Knowledge base	Test set[a]	Correctly predicted/observed hydration Conserved	Displaced	Predictive accuracy	k	C_m	Feature weights[b] (Density, Hydrophilicity, B-value, H-bonds)
Training for active-site optimization (overlap between knowledge base and test set, some bias[c])							
118 active-site waters (59 cons., 59 disp.)	157 active-site waters	37/42	79/115	73.9%	3	0.504	0.530, 0.047, 0.394, 0.029
1700 first-shell (850 cons., 850 disp.)	**157 active-site**	**35/42**	**86/115**	**77.1%**	**3**	**0.524**	**0.107, 0.334, 0.269, 0.290**
1700 first-shell	157 active-site	29/42	82/115	70.7%	7	0.365	0.324, 0.001, 0.438, 0.237
2832 first-shell (1416 cons., 1416 disp.)	**224 active-site**	**48/59**	**130/165**	**79.5%**	**3**	**0.549**	**0.512, 0.006, 0.229, 0.253**
Training for first-shell optimization (overlap between knowledge base and test set, but no self-vote[c])							
1700 first-shell	1700 first-shell	588/850	509/850	64.5%	7	0.292	0.212, 0.333, 0.040, 0.415
1700 first-shell	**1700 first-shell**	**619/850**	**534/850**	**67.8%**	**39**	**0.356**	**0.026, 0.254, 0.237, 0.483**
2832 first-shell	2832 first-shell	992/1416	931/1416	67.9%	7	0.369	0.023, 0.295, 0.341, 0.341[d]
2832 first-shell	2832 first-shell	1032/1416	905/1416	68.4%	39	0.369	0.234, 0.170, 0.234, 0.362[d]
2832 first-shell	**2832 first-shell**	**1045/1416**	**902/1416**	**68.8%**	**39**	**0.377**	**0.244, 0.175, 0.240, 0.341**
Unbiased knn testing for active-site prediction (no overlap between knowledge base and test set)							
1700 first-shell	**67 active-site**	**11/16**	**39/51**	**74.6%**	**3**	**0.406**	**from 1700/157 k = 3 knnGA**
1700 first-shell	67 active-site	6/16	37/51	64.2%	7	0.094	from 1700/157 k = 7 knnGA
Unbiased knn testing for first-shell prediction (no overlap between knowledge base and test set)							
1700 first-shell	1545 first-shell	539/905	401/640	60.8%	7	0.219	from 1700/1700 k = 7 knnGA
1700 first-shell	**1545 first-shell**	**523/905**	**419/640**	**61.0%**	**39**	**0.229**	**from 1700/1700 k = 39 knnGA**

Boldface indicates runs discussed in detail in the text.

[a] The 157, 67, and 224 test sets are, respectively, all active-site waters in the 13 proteins in Table 1, in the 7 proteins in Table 2, and in the 20 proteins in Tables 1 and 2. The 1700, 1545, and 2832 test sets are, respectively, the largest balanced set of conserved and displaced first-shell waters in the proteins in Table 1, all first-shell waters in the proteins in Table 2, and the largest balanced set of conserved and displaced waters in the proteins in Tables 1 and 2.

[b] Weights for each run have been normalized to sum to 1.

[c] Bias arises when there is overlap between the knowledge base and test set. We have minimized bias in the training of the knnGA by disallowing self-votes in the knn when the training and test sets are the same; thus, a water cannot vote on its own conserved or displaced status, and its classification depends entirely on the environmental chemistry of other water molecules.

[d] Indicates runs in which a parsimony constant of 0.04 was introduced; for all other runs the parsimony constant was 0.

overlap introduced some bias, since each water molecule included in both the knowledge base and test set would correctly vote on its own status, constituting a ''self-vote'' in the knn procedure. However, for $k = 3$, two other water molecules also contributed to the predicted status. (For later training on first-shell sites and knn prediction on active-site waters, self-votes were disallowed in the algorithm.) The prediction results from these initial tests (top of Table 3) were encouraging: *Consolv* was able to correctly predict conserved/displaced status for 121 out of 157, or 77.1%, of the active-site waters from the thirteen structures in Table 1. The C_m value of 0.52 for this test indicates good predictive accuracy and good balance, since 75% accuracy for both the conserved and displaced classes yields a C_m value of 0.50; for perfect prediction, $C_m = 1$, for 50% correct prediction in both classes, $C_m = 0$, and for entirely incorrect predictions, $C_m = -1$. When *Consolv* was trained solely on a knowledge base consisting of a balanced set of 59 conserved and 59 displaced active-site water molecules from all structures, rather than on the balanced set of 1700 first-shell waters, the prediction rate for the 157 active-site water molecules was 3.2% lower (73.9% correct; Table 3). This is probably due to the smaller sample size (118) of balanced conserved and displaced active-site water molecules not providing as complete a characterization of favored water environments. Conversely, when the knowledge base for training was expanded to include the largest balanced set of 1416 conserved and 1416 displaced water sites from the 20 proteins in Tables 1 and 2, the predictive accuracy increased to 79.5% for the 224 active-site waters in these proteins. A genetic program version of *Consolv*, which allows non-linear scaling of feature axes, provided 79.0% accuracy when trained on the 1700-water knowledge base to predict the conservation of 157 active-site water molecules (Raymer *et al*., 1996).

To assess whether feature weights were linked, and, therefore, cooptimized on the GA chromosome during training, runs were executed for the 1700-water knowledge base and the 157 active-site water test set with $k = 3$ for three different chromosome orderings: {adn, ahp, bval, hbd}, {ahp, hbd, adn, bval}, and {hbd, bval, adn, ahp}, where adn = atomic density, ahp = atomic hydrophilicity, bval = temperature factor, and hbd = number of hydrogen bonds. Each run achieved the same predictive accuracy (77.1%) independent of the chromosome ordering, indicating that linkage is not a problem.

For prediction of first-shell hydration conserved between independently determined structures, *Consolv* was trained on the 1700-water and 2832-water knowledge bases (middle of Table 3). Maximum predictive accuracy, 68.8% ($C_m = 0.38$), was obtained for $k = 39$ with the 2832-water knowledge base. Including a parsimony term in the GA fitness function to minimize the magnitude of individual weights reduced the accuracy only 0.4% for an otherwise identical run. Somewhat lower accuracy was found for $k = 7$ first-shell training using the 1700-water (64.5%) and the 2832-water (67.9%) knowledge bases. Training results for first-shell ($k = 7$ and $k = 39$) and active-site waters ($k = 3$) showed that use of the larger, 2832-water knowledge base improved accuracy from 1 to 3.4%. It was unexpected that training on conservation of first-shell hydration would have lower accuracy than training on conservation of active-site water molecules upon ligand binding. A possible explanation is that active-site bound water molecules are more likely to be carefully assigned and refined by crystallographers due to their functional importance; therefore the *Consolv* training data may have contained fewer missing or spurious water assignments in active sites than were found in the first hydration shells.

Analysis of water binding determinants

Each *Consolv* training session resulted in a weight set specifying the relative importance of the water molecule's temperature factor and the hydrogen-bonding potential, atomic density, and atomic hydrophilicity of the water molecule's neighborhood in determining whether the water was conserved or displaced. However, because some features were highly correlated, as shown in Table 4, several different *Consolv* weight sets could give similarly accurate predictions. For instance, the site's atomic hydrophilicity was measured essentially as hydrophilicity-weighted atomic density, giving a high correlation coefficient with atomic density (0.64). The number of hydrogen bonds correlated both with the density of atomic neighbors (0.40) and their hydrophilicity (0.78), due to the potentially greater availability of hydrogen-bond partners. Temperature factor had a relatively low anti-correlation (between –0.28 and –0.32) with the three other environmental features. Analysis of the feature weights for the knnGA training runs in Table 3 indicated that one of two highly correlated features, atomic density and atomic hydrophilicity, was the most important discriminator between conserved and displaced active-site water molecules in each of the three most accurate training runs; when one of these two features had a large

Table 4. Pearson product–moment correlation between environmental features for first-shell waters from 20 structures

	Atomic density	Atomic hydrophilicity	Temperature factor	Number of hydrogen bonds
Atomic density	1.000			
Atomic hydrophilicity	0.642	1.000		
Temperature factor	–0.292	–0.318	1.000	
Number of hydrogen bonds	0.405	0.783	–0.278	1.000

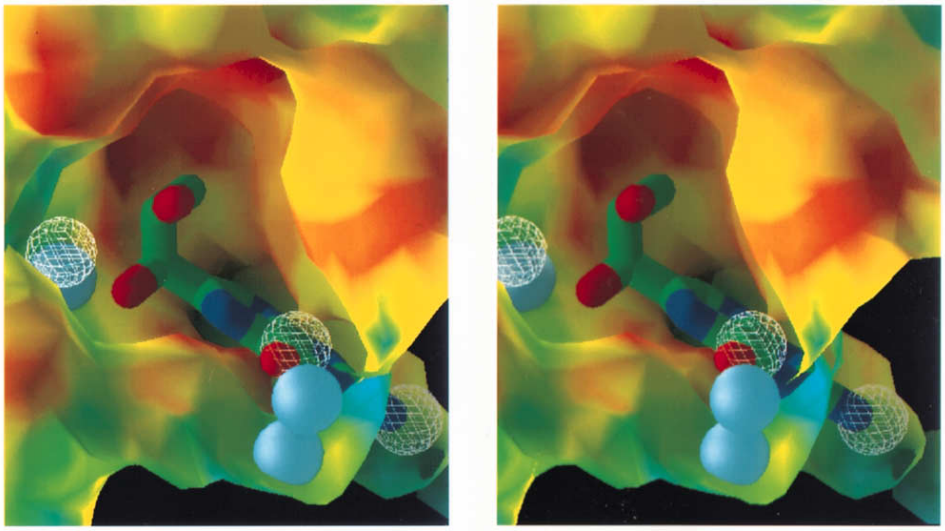


Figure 7. Stereo view of water-mediated and polar ligand atom interactions in the active site of dihydrofolate reductase. The solvent-accessible molecular surface of ligand-free dihydrofolate reductase (DHFR; PDB structure 1DR2, McTigue *et al*., 1992) is colored by the sum of the atomic hydrophilicity values of protein atoms and water molecules within 3.6 Å of the surface, allowing analysis of the contributions of hydrophilicity and hydrophobicity to water and biopterin binding. Biopterin (tubes with carbon atoms colored green, oxygen colored red, and nitrogen blue; from PDB structure 1DR3; McTigue *et al*., 1992) is positioned based on main-chain superimposition between the ligand-bound and free structures (RMSD = 0.12 Å). The six active-site water molecules from the ligand-free structure are shown as spheres; blue spheres are water molecules conserved in the complex, while mesh spheres are waters displaced upon biopterin binding. The displaced water at center was supplanted by an oxygen atom in biopterin (mesh sphere surrounding red tube). The water at right was displaced by a nitrogen atom in biopterin (mesh surrounding blue tube), and is an example of a water molecule predicted by *Consolv* to be conserved, but actually substituted by a similarly polar ligand atom. Atomic hydrophilicity values are: 0.08 hydrations per carbon or sulfur atom, 0.35 per neutral nitrogen, 0.44 per positively charged nitrogen, 0.51 per negatively charged oxygen, and 0.53 per neutral oxygen atom (Kuhn *et al*., 1995). Surface colors range from red, most hydrophobic (atomic hydrophilicity of 0.1, or $\sim$1 carbon or sulfur neighbor), to yellow (atomic hydrophilicity of 1.5) and green (atomic hydrophilicity of 3.0), to blue, most hydrophilic (atomic hydrophilicity >4, equivalent to 8 or more hydrophilic neighbors). The solvent-accessible protein surface was calculated using a 1.4 Å radius probe sphere and the following van der Waals radii including implicit hydrogens: O, 1.40 Å; OH, 1.60 Å; N, 1.54 Å; NH, 1.70 Å; NH_2, 1.80 Å; NH_3, 2.00 Å; CH, CH_2, CH_3, 2.00 Å; C, 1.74 Å; CH(sp^2), 1.86 Å; S, 1.80 Å; and SH, 1.85 Å. The surface was generated as a triangulated surface using *MSP* (Connolly, 1993; http://www.biohedron.com) and visualized using *AVS* (Upson *et al*., 1989; Advanced Visual Systems, Inc., Waltham, MA) using modules developed by Michael Pique and colleagues at The Scripps Research Institute.

weight, the other had a small weight, consistent with the second feature not contributing much additional information for classification.

To identify a consistent and parsimonious (minimal magnitude) weight set, 20 independent knnGA runs were performed with the 1700-water knowledge base and 157-water test set, different initial random seeds, $k = 3$, and a 150-fold range in parsimony constants (0.002 to 0.3). These runs had an average accuracy of 76.4% for active-site water classification, as compared to 77.1% without parsimony; the accuracy of the worst run was 73.9%, and 15 of the runs achieved 77.1% accuracy. (For brevity, these 20 runs are not listed in Table 3.) Furthermore, when the weights for each run were normalized to sum to 1, the mean and standard deviation in weights over these parsimony runs were consistent with earlier results: an average atomic density weight of 0.11 (std. dev. 0.0076); atomic hydrophilicity, 0.37 (0.059); temperature factor, 0.26 (0.054); and number of hydrogen bonds, 0.25 (0.022). In the most accurate active-site training run (2832-water knowledge base/224-water test set with $k = 3$, 81.4% accuracy on conserved water prediction, 78.8% accuracy on displaced water prediction, and 79.5% overall accuracy), atomic density was the most important feature for classifying active-site waters, with temperature factor and number of hydrogen bonds each contributing approximately one-half as much. The importance of atomic density for classifying active-site waters may reflect the propensity of water to bind in surface grooves, since atomic density is related to groove depth (Kuhn *et al*., 1992). For first-shell training, the number of hydrogen bonds was consistently the most important feature for discriminating between conserved and non-conserved waters in all training runs (1700/1700 at $k = 7$ and $k = 39$, and 2832/2832 at $k = 7$ with parsimony and at $k = 39$ with and without parsimony; middle of Table 3).

Unbiased predictions on new structures

After tuning the search parameters and performing self-consistency tests on water molecules in the

knowledge base, *Consolv's* weighted knn algorithm was applied to active-site water molecules from seven new, non-homologous protein structures (Table 2). No information about the new structures was provided during the training phase, so results for these structures provide an unbiased indication of predictive ability. These predictions ran much more quickly because the knnGA was not being trained on the new structures. Rather, the optimized weight set produced by the genetic algorithm during previous training (e.g. on the 157 active-site waters with the 1700 first-shell water knowledge base) was used by the weighted knn classifier alone to predict water status in the new structures. KnnGA training runs on a large test set (1700-water knowledge base and 1700-water test set) typically took 12 hours elapsed time (reduced from four days by implementing the branch and bound knn algorithm (bbknn) of Fukunaga & Narendra, 1975) running on one processor of a 50 MHz SPARCstation (Sun SPARC20-502). Predictions on hundreds of new water sites using the stand-alone weighted bbknn took less than a second elapsed time. This will allow a fast, portable version of *Consolv* to be provided for other laboratories' use, consisting of the knowledge base, optimal k value and weight set (the $k = 39$ 2832/2832 weights optimized for first-shell prediction and the $k = 3$ 2832/224 weights optimized for active-site prediction), software for calculating water environments in new proteins, and the weighted bbknn classifier module of *Consolv*.

For the 67 active-site waters in the seven new proteins, *Consolv* achieved an unbiased predictive accuracy of 74.6% (50/67 correct predictions, $C_m = 0.41$; bottom of Table 3) using a weight set derived from training on the 1700-water knowledge base with $k = 3$. For conserved water molecules the predictive accuracy was 68.8%, while for displaced waters the accuracy was 76.5%. Training and

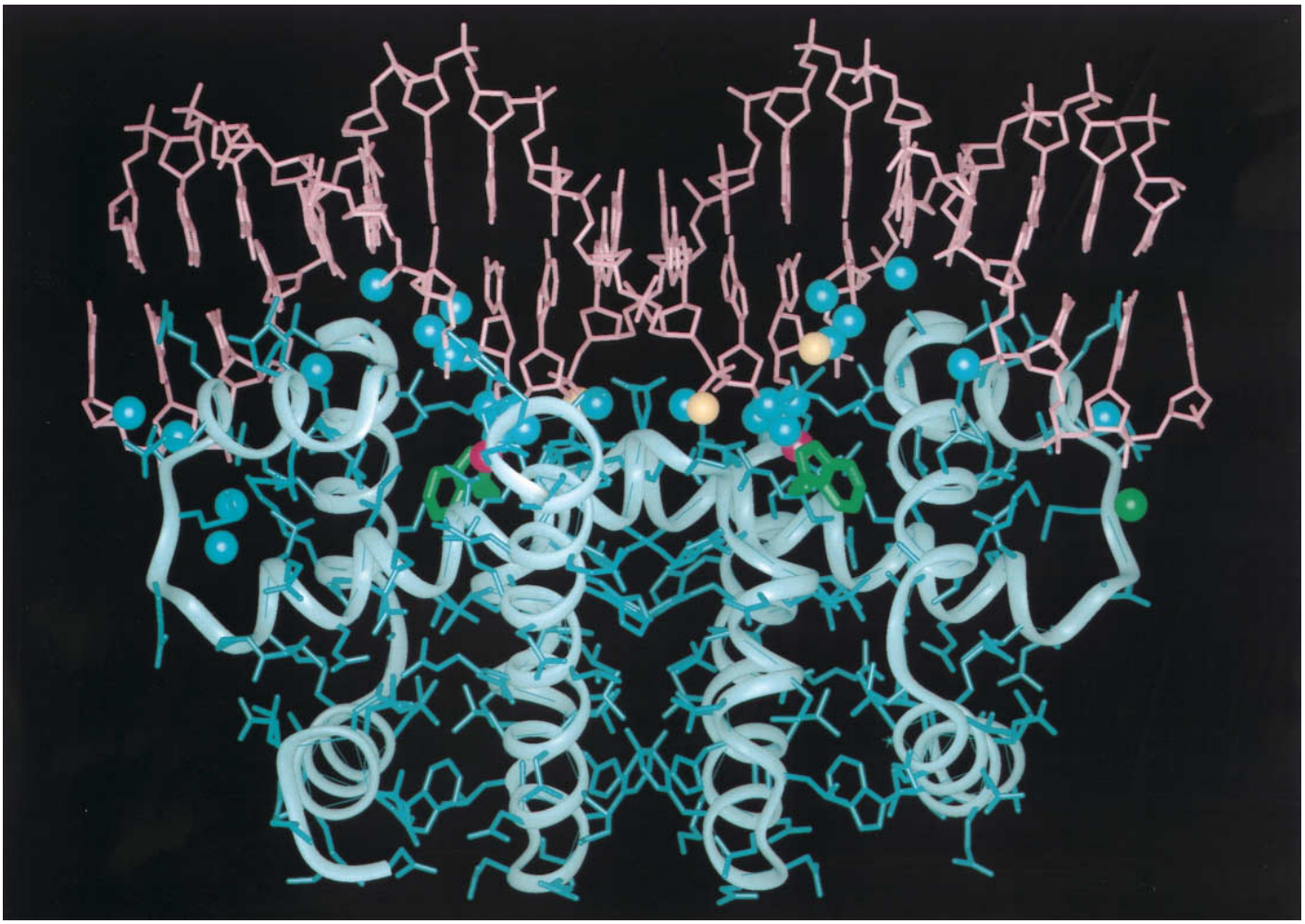


Figure 8. Hydration of the Trp repressor. Spheres shown at the protein-DNA interface are water molecules present in the ligand-free Trp repressor structure (blue ribbons and tubes; PDB structure 2WRP; Zhang *et al*., 1987). The DNA ligand (pink tubes; from PDB structure 1TRO; Joachimiak *et al*., 1994; Otwinowski *et al*., 1988) is shown for reference, and positioned based on main-chain superposition of the ligand-bound repressor onto the free repressor. Water molecules shown as blue spheres were correctly predicted by *Consolv* to be displaced in the ligand-bound structure, and water molecules shown in magenta were correctly predicted to be conserved. Peach-colored water molecules were predicted to be displaced but were actually conserved, whereas the green water molecule at far right was predicted to be conserved, but was actually displaced in the ligand-bound structure.

testing results support the use of small k values, particularly $k = 3$, for active-site prediction; results for $k = 7$, which have moderate accuracy, are included in Table 3 to facilitate comparison between active-site and first-shell training and testing runs. *Consolv's* unbiased $k = 3$ predictive accuracy on the proteins from Table 2 (1CGT, 75%; 1DR2, 67%; 2CTV, 71%; 2WRP, 88%; 3ENL, 60%; RTEM, 50%; and GST, 33%; with 74.6% overall accuracy for their 67 active-site waters) correlated only with the number of active-site bound waters in the free structure, not with the structure's resolution or *R*-factor, nor those of the ligand-bound structure used to check the predictions. Using the 1700-water knowledge base, there was a modest 2.5% decrease in predictive accuracy between biased prediction on the 157 active-site waters and unbiased prediction on the 67 new active-site waters. *Consolv* was also used for unbiased prediction of first-shell water molecule conservation in the seven new protein pairs (bottom of Table 3). Similar predictive accuracy and balance for these 1545 water sites was attained using a weight set derived from $k = 7$ knnGA training (60.8%, $C_m = 0.22$) and from $k = 39$ training (61.0%, $C_m = 0.23$).

Water conservation in multiple independently-solved structures

A related application for *Consolv* is to evaluate the likelihood of conserved hydration in several independent structures of a protein with the same ligand-binding status. To evaluate its suitability for this purpose, we applied *Consolv's* weighted knn to the water sites in each of three high-resolution bovine β-trypsin structures (PDB 1TPO, 2PTN, and 3PTN) using $k = 39$ and weights derived for first-shell prediction with the 2832-water knowledge base. Using weights without parsimony, *Consolv's* accuracies on predicting conservation of water sites in 1TPO, 2PTN, and 3PTN relative to the bovine pancreatic trypsin inhibitor complex 2PTC, were, respectively, 66.2%, 69.7%, and 70.5% ($C_m = 0.42$, 0.44, and 0.48); with weights derived using a parsimony constant of 0.04, the results were 64.9%, 71.0%, and 70.5% ($C_m = 0.40$, 0.46, and 0.49). These tests had an average of 8% greater accuracy on predicting first-shell conserved water for trypsin structures than was found for the seven structural pairs in earlier unbiased tests (Table 3). This may be attributable to the use of a larger knowledge base (2832 waters *versus* 1700) or to the accurate crystallographic assignment of water sites in the four high-resolution (1.6 to 1.9 Å) trypsin structures.

Testing other statistical methods for classifying water sites

An unweighted knn classifier, trained using the 1700-water *Consolv* knowledge base and the 157 active-site water test set with $k = 3$, achieved a prediction rate of 66.9%, about 10% lower than *Consolv's* weighted knn result, 77.1%. Discriminant analysis on the same data using the *Discrim* function in SAS software resulted in prediction rates of 49.7% for a linear discriminant function, and 51.6% for a quadratic function. This random level of prediction is not surprising, since parametric discriminant analysis is more appropriate for classifying data with well-separated Gaussian distributions in the multi-feature space.

Threshold-based decision rules were also compared with *Consolv*. Assessing the proportion of water sites satisfying a feature threshold criterion (e.g. *B*-value >70 Å^2) that are conserved or displaced gave similar results for all four environmental features. The feature threshold was finely sampled, and when the threshold was set such that a reasonable number of water molecules qualified, no useful tendency towards conservation or displacement could be detected. Only when the threshold was set to an extreme value, resulting in a small number of waters satisfying the criterion, was there a strong tendency towards conservation or displacement. For example, of 16 water molecules in the knowledge base with *B*-value >80 Å^2, only one was conserved, but these 16 waters represented only 0.4% of the waters in the knowledge base; when the threshold was set to *B*-value >70 Å^2, fewer than 3% of the knowledge-base waters were included, but the proportion of conserved waters was already 37%. Thus, threshold-based rules are not a practical means for predicting water site conservation.

Displacement of active-site bound water by polar ligand atoms

Examination of active-site water molecules mispredicted by *Consolv* revealed several interesting trends. 71% of mispredictions in the seven new structures involved displaced waters that *Consolv* predicted to be conserved. Of the 12 displaced water molecules predicted to be conserved, 8 were displaced by an oxygen atom and 2 by a nitrogen atom from the ligand; thus, 83% of water sites mispredicted to be conserved were replaced by a polar ligand atom. *Consolv* was therefore correct in determining that the micro-environment of these sites favored conserved polar atom binding. When the definition of conserved sites included those occupied by water or a polar ligand atom in the ligand-bound structure (sites of ''effective solvation''; Zhang & Matthews, 1994; Faerman & Karplus, 1995), *Consolv's* accuracy increased to 89.7% ($C_m = 0.78$), indicating high accuracy for detecting conserved polar atom binding. In dihydrofolate reductase, for example, one of two mispredicted waters was predicted to be conserved, but was actually displaced by a polar ligand atom. Figure 7 shows the active-site water molecules from the ligand-free structure of dihydrofolate reductase with actual conserved or displaced status indicated

respectively by solid or mesh spheres, the ligand-free surface colored by atomic hydrophilicity, and the ligand, biopterin, superimposed from the structure of the complex (McTigue *et al*., 1992). The water site at extreme right (white mesh surrounding blue tube) was predicted by *Consolv* to be conserved and was actually displaced by a nitrogen atom in biopterin.

Ligand hydration

The importance of ligand hydration for some complexes is suggested by *Consolv's* predictions on the Trp repressor (Figure 8). For this protein, nearly all water molecules in the free structure were correctly predicted to be displaced upon ligand binding. However, many water molecules do mediate the protein–DNA interface of the Trp repressor complex (Otwinowski *et al*., 1988), and free nucleic acid structures are typically well hydrated (Berman, 1994; Westhof, 1988), suggesting that DNA-bound hydration is involved. For ligands with structures solved independently of the protein, the contribution of ligand hydration to complex formation can be evaluated by the same approaches used here for proteins: assessing water conservation among superimposed, independently-solved structures of the ligand, and analyzing the ligand micro-environment of bound water molecules using *Consolv*.

Consolv enhancements and applications

An intrinsic feature of *Consolv* is the ability to train on additional protein structures and to test other environmental features of water molecules' environments for their ability to improve discrimination between conserved and displaced water sites. Examples of such features are electrostatic potential (Gilson & Honig, 1988); thermal mobility measured by an index incorporating both temperature factor and occupancy (L. A. Kuhn and P. C. Sanschagrin, unpublished results) evaluated separately for the water molecule and for neighboring protein atoms; separate counts of main-chain and side-chain hydrogen bonds and hydrogen bonds to other water molecules; and evolutionary sequence conservation in the neighborhood of the water site, determined from multiple-sequence alignments. Superimposed, independently solved structures can be used to define more reliable, consensus water sites (Faerman & Karplus, 1995) and quantify the degree of water site conservation for training *Consolv*, removing the restriction that the knowledge base include only those proteins with structures known in both the ligand-bound and free state. This greatly increases the number of high-resolution ($\leqslant$2.0 Å) structures available for training and also allows prediction of the relative likelihood of water site conservation, based on the observed degree of conservation of environmentally similar waters.

A key goal is to apply *Consolv* prediction to improve protein structure-based inhibitor design. Appropriately incorporating conserved water has been a missing link in the analysis of protein recognition, and *Consolv* provides a means for modeling active-site water conservation with 75% accuracy. This unbiased accuracy for two-state (conserved, displaced) classification of water sites is comparable to the best of current three-state (helix, strand, loop) secondary structure prediction methods, 70 to 72% (Mehta *et al*., 1995; Rost & Sander, 1993). *Consolv's* ability to predict conservation of water or polar ligand atom binding with 90% accuracy, independent of the ligand, provides a rational basis for refining drug design templates to appropriately include sites likely to conserve water or bind a polar ligand atom, as well as to disinclude water molecules likely to be displaced. Protein mutagenesis, drug design, and molecular simulations will benefit from a more complete representation of proteins that includes conserved bound water.

The *Consolv* knowledge base and knn water classification software will be made available to other researchers; please contact Michael Raymer or Leslie Kuhn (*raymermi@sol.bch.msu.edu; kuhn@agua.bch.msu.edu*). For related information, see *http://www.bch.msu.edu/labs/kuhn/web/index.html*.

Acknowledgements

The authors thank Drs Eric Torng and Anil Jain (Michigan State University) for advice on algorithm optimization and statistical pattern recognition, Drs John Tainer (The Scripps Research Institute) and Thomas Stout (University of California, San Francisco) for critical feedback on the manuscript, and Michael Pique (The Scripps Research Institute) for providing *AVS* modules and guidance. L.K. dedicates this paper to the memory of Dr Jairo Arevalo, a wonderful scientist and person. This research was supported by the Research Excellence Fund for Computing and Technology, the Research Excellence Fund for Protein Structure, Function, and Design, an All-University Research Initiation Grant, and by National Science Foundation Grant BIR9600831 to L.K.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein Data Bank. In *Crystallographic Databases–Information Content, Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G. & Sievers, R., eds), pp. 107–132. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Badger, J. (1993). Multiple hydration layers in cubic insulin crystals. *Biophys. J.* **65**, 1656–1659.

Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.

Berman, H. M. (1994). Hydration of DNA: take 2. *Curr. Opin. Struct. Biol.* **4**, 345–350.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O.,

Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
Connolly, M. L. (1993). The molecular surface package. *J. Mol. Graphics,* **11**, 139–141.
Dandekar, T. & Argos, P. (1994). Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.* **236**, 844–861.
Edsall, J. T. & McKenzie, H. A. (1983). Water and proteins. II. The location and dynamics of water in protein systems and its relation to their stability and properties. *Advan. Biophys.* **16**, 53–183.
Faerman, C. H. & Karplus, P. A. (1995). Consensus preferred hydration sites in six FKBP12-drug complexes. *Proteins: Struct. Funct. Genet.* **23**, 1–11.
Fauman, E. B., Rutenber, E. E., Maley, G. F., Maley, F. & Stroud, R. M. (1994). Water-mediated substrate/product discrimination: the product complex of thymidylate synthase at 1.83 Å. *Biochemistry,* **33**, 1502–1511.
Fitzpatrick, P. A., Steinmetz, A. C. U., Ringe, D. & Klibanov, A. M. (1993). Enzyme crystal structure in a neat organic solvent. *Proc. Natl Acad. Sci. USA,* **90**, 8653–8657.
Fukunaga, K. & Narendra, P. M. (1975). A branch and bound algorithm for computing *k*-nearest neighbors. *IEEE Transactions on Computers*. July, 750–753.
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.
Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, San Mateo, CA.
Goodfellow, J. M. & Vovelle, F. (1989). Biomolecular energy calculations using transputer technology. *Eur Biophys. J.* **17**, 167–172.
Gros, P., Teplyakov, A. V. & Hol, W. G. J. (1992). Effects of eglin-c binding to thermitase: three-dimensional structure comparison of native thermitase and thermitase eglin-c complexes. *Proteins: Struct. Funct. Genet.* **12**, 63–74.
Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.
Hutchinson, E. G. & Thornton, J. M. (1990). HERA: a program to draw schematic diagrams of protein secondary structures. *Proteins: Struct. Funct. Genet.* **8**, 203–212.
Jiang, J.-S. & Brünger, A. T. (1994). Protein hydration observed by X-ray diffraction: solvation properties of penicillopepsin and neuraminidase crystal structures. *J. Mol. Biol.* **243**, 100–115.
Joachimiak, A., Haran, T. E. & Sigler, P. B. (1994). Mutagenesis supports water mediated recognition in the Trp repressor-operator system. *EMBO J.* **13**, 367–372.
Jones, G., Willett, P. & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245**, 43–53.
Karplus, P. A. & Faerman, C. (1994). Ordered water in macromolecular structure. *Curr. Opin. Struct. Biol.* **4**, 770–776.
Kelly, J. D. & Davis, L. (1991). Hybridizing the genetic algorithm and the *k* nearest neighbors classification algorithm. *Proceedings of the Fourth International Conference on Genetic Algorithms and their Applications*, pp. 377–383.
Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Tainer, J. A. (1992). The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *J. Mol. Biol.* **228**, 13–22.
Kuhn, L. A., Swanson, C. A., Pique, M. E., Tainer, J. A. & Getzoff, E. D. (1995). Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins: Struct. Funct. Genet.* **23**, 536–547.
Kuntz, I. D. & Kauzmann, W. (1974). Hydration of proteins and polypeptides. *Advan. Protein Chem.* **28**, 239–345.
Lam, P. Y. S., Jadhav, P. K., Eyermann, C. J., Hodge, C. N., Ru, Y., Bacheler, L. T., Meek, J. L., Otto, M. J., Rayner, M. M., Wong, Y. N., Chang, C.-H., Weber, P. C., Jackson, D. A., Sharpe, T. R. & Erickson-Viitanen, S. (1994). Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science,* **263**, 380–384.
Le Grand, S. M. & Merz, K. M., Jr (1994). The genetic algorithm and protein tertiary structure prediction. In *The Protein Folding Problem and Tertiary Structure Prediction* (Merz, K. M., Jr & Le Grand, S. M., eds), pp. 109–124, Birkhäuser, Boston.
McTigue, M. A., Davies, J. F., Kaufman, B. T. & Kraut, J. (1992). Crystal structure of chicken liver dihydrofolate reductase complexed with $NADP^+$ and biopterin. *Biochemistry,* **31**, 7264–7273.
McTigue, M. A., Williams, D. R. & Tainer, J. A. (1995). Crystal structures of a schistosomal drug and vaccine target: glutathione *S*-transferase from *Schistosoma japonica* and its complex with the leading antischistosomal drug praziquantel. *J. Mol. Biol.* **246**, 21–27.
Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta,* **405**, 442–451.
May, A. C. W. & Johnson, M. S. (1994). Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Protein Eng.* **7**, 475–485.
Mehta, P. K., Heringa, J. & Argos, P. (1995). A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci.* **4**, 2517–2525.
Oshiro, C. M., Kuntz, I. D. & Dixon, J. S. (1995). Flexible ligand docking using a genetic algorithm. *J. Comp. Aided Mol. Design,* **9**, 113–130.
Otwinowski, Z., Schevitz, R. W., Zhang, R.-G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. & Sigler, P. B. (1988). Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature,* **335**, 321–329.
Overington, J. P., Johnson, M. S., Sali, A. & Blundell, T. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues, and structure prediction. *Proc. Roy. Soc. Lond. ser. B*, **241**, 132–145.
Patton, A. L., Punch, W. F. & Goodman, E. D. (1995). A standard GA approach to native protein conformation prediction. *Proceedings of the International Conference on Genetic Algorithms*, Pittsburgh, PA.
Pitt, W. R., Murray-Rust, J. & Goodfellow, J. M. (1993). AQUARIUS2: Knowledge-based modelling of solvent sites around proteins. *J. Comp. Chem.* **14**, 1007–1018.
Punch, W. F., Goodman, E. D., Pei, M., Chia-Shun, L., Hovland, P. & Enbody, R. (1993). Further research on feature selection and classification using genetic

algorithms. *Proceedings of the International Conference on Genetic Algorithms*, pp. 557–564.
Raymer, M. L., Punch, W. F., Goodman, E. D. & Kuhn, L. A. (1996). Genetic programming for improved data mining–application to the biochemistry of protein interactions. In *Genetic Programming 96: Proceedings of the First Annual Conference* (Koza, J. R., Goldberg, D. E., Fogel, D. B. & Riolo, R. L., eds), pp. 375–381, MIT Press, Cambridge, Massachusetts.
Ring, C. S. & Cohen, F. E. (1994). Conformation sampling of loop structures using genetic algorithms. *Isr. J. Chem.* **34**, 245–252.
Roe, S. M. & Teeter, M. M. (1993). Patterns for prediction of hydration around polar residues in proteins. *J. Mol. Biol.* **229**, 419–427.
Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.
Salamov, A. A. & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* **247**, 11–15.
Schraudolph, N. N. & Grefenstette, J. J. (1992). A user's guide to GAUCSD. Computer Science Department, University of California, San Diego, CA, Technical Report CS92-249.
Siedlecki, W. & Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters,* **10**, 335–347.
Strynadka, N. C. J., Adachi, H., Jensen, S. E., Johns, K., Sielecki, A., Betzel, C., Sutoh, K. & James, M. N. G. (1992). Molecular structure of the acyl-enzyme intermediate in β-lactam hydrolysis at 1.7 Å resolution. *Nature,* **359**, 700–705.
Tanford, C. (1980). *The Hydrophobic Effect*. Wiley-Interscience, New York.
Travis, J. (1993). Proteins and organic solvents make an eye-opening mix. *Science,* **262**, 1374.
Unger, R. & Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.* **231**, 75–81.
Upson, C., Faulhaber, T., Jr, Kamins, D., Laidlaw, D., Schlegel, D., Vroom, J., Gurwitz, R. & van Dam, A. (1989). The application visualization system: a computational environment for scientific visualization. *IEEE Comp. Graphics Appl.* **9**, 30–42.
Vedani, A. & Huhta, D. W. (1991). An algorithm for the systematic solvation of protein based on the directionality of hydrogen bonds. *J. Am. Chem. Soc.* **113**, 5860–5862.
Wade, R. C., Bohr, H. & Wolynes, P. G. (1992). Prediction of water binding sites by neural networks. *J. Am. Chem. Soc.* **114**, 8284–8285.
Wade, R. C., Clark, K. J. & Goodford, P. J. (1993). Further developments of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* **36**, 140–147.
Walters, D. E. & Hinds, R. M. (1994). Genetically evolved receptor models: a computational approach to construction of receptor models. *J. Med. Chem.* **37**, 2527–2536.
Westhof, E. (1988). Water: an integral part of nucleic acid structure. *Annu. Rev. Biophys. Biophys. Chem.* **17**, 125–144.
Wilson, I. A. & Fremont, D. H. (1993). Structural analysis of MHC Class I molecules with bound peptide antigens. *Semin. Immunol.* **5**, 75–80.
Wlodawer, A., Miller, M., Jaskólski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J. & Kent, S. B. H. (1989). Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science,* **245**, 616–621.
Yi, T.-M. & Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.* **232**, 1117–1129.
Zhang, L. & Hermans, J. (1996). Hydrophilicity of cavities in proteins. *Proteins: Struct. Funct. Genet.* **24**, 433–438.
Zhang, R.-G., Joachimiak, A., Lawson, C. L., Schevitz, R. W., Otwinowski, Z. & Sigler, P. B. (1987). The crystal structure of Trp aporepressor at 1.8 angstroms shows how binding tryptophan enhances DNA affinity. *Nature,* **327**, 591–597.
Zhang, X.-J. & Matthews, B. W. (1994). Conservation of solvent-binding sites in 10 crystal forms of T4 lysozyme. *Protein Sci.* **3**, 1031–1039.

Edited by B. Honig

(Received 27 February 1996; received in revised form 26 September 1996; accepted 1 October 1996)