J Mol Model (2003) 9:172–182
DOI 10.1007/s00894-003-0129-x

ORIGINAL PAPER

Alfonso T. García-Sosa · Ricardo L. Mancera · Philip M. Dean

WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes

Received: 20 November 2002 / Accepted: 5 March 2003 / Published online: 17 May 2003
© Springer-Verlag 2003

Abstract We have performed a multivariate logistic regression analysis to establish a statistical correlation between the structural properties of water molecules in the binding site of a free protein crystal structure, with the probability of observing the water molecules in the same location in the crystal structure of the ligand-complexed form. The temperature B-factor, the solvent-contact surface area, the total hydrogen bond energy and the number of protein–water contacts were found to discriminate between bound and displaceable water molecules in the best regression functions obtained. These functions may be used to identify those bound water molecules that should be included in structure-based drug design and ligand docking algorithms.

Keywords Protein hydration · Drug design · Bound water molecules · Multivariate logistic regression

Introduction

Considerable effort has historically and recently been directed towards understanding and predicting the water molecules observed in protein–ligand complexes. This introduction mentions a few approaches to biomolecular hydration, both experimental and computational. We present the case for our use of statistical methods of analyses of collections of protein structures in order to be able to compare with other methods and show how this study and the WaterScore program work reasonably well in a wide variety of biomolecular systems and hydration conditions.

Protein crystal structures reveal that water molecules often engage in hydrogen bonding with acceptor and donor groups on the protein surface, and sometimes also with other water molecules which themselves may form hydrogen bonds to the protein. [1] The latter are often called the first hydration shell of a protein. Water molecules may also interact with both the protein and a bound ligand, mediating polar interactions between them. These interactions can help to form complex hydrogen-bond networks that are further stabilized through cooperativity effects [2] and can have a role in biomolecular structure, function, recognition and specificity. [3, 4, 5, 6]

It has been observed that certain water molecules occupy the same positions in crystal structures of the same protein under different crystallographic conditions, [7] and/or with different ligands, [8, 9, 10] or in a set of structurally related proteins. [4, 11, 12, 13, 14] Furthermore, water molecules found on protein surfaces and mediating interactions with ligands have been observed to have characteristic structural properties, such as binding in grooves on the protein surface [15] and making on average three hydrogen bonds with the protein. [16]

Crystal packing effects might be responsible for overstabilizing a protein's hydration structure compared to the physiological aqueous environment, [17] or for changing the location of surface water molecules. [12, 18] The determination of conserved water sites in X-ray structure determinations also varies as the environment of the protein changes (temperature, solvent conditions, pH, hydration level). [19] However, as the resolution of protein crystal structures improves, more and better-defined structured water sites are being found, [20] as had been predicted earlier. [21] It has been suggested that the hydration sites observed in sets of multiple crystal structures of a protein actually represent the configurational space sampled by water molecules of that protein, with the equilibrium between different hydration sites shifting according to the environmental conditions. [19]

A. T. García-Sosa (✉)
Department of Pharmacology,
University of Cambridge,
Tennis Court Road, Cambridge, CB2 1PD, UK
e-mail: atg21@cam.ac.uk
Tel.: +44 1223 238042, Fax: +44 1223 238088

R. L. Mancera · P. M. Dean
De Novo Pharmaceuticals,
Compass House, Vision Park, Chivers Way, Histon,
Cambridge, CB4 9ZR, UK

Crystallographic water molecules also seem to be major contributors to the energetics of protein–ligand and protein–protein complexes. In antigen–antibody complexes, for example, water molecules are seen to provide an energy-regulatory function by means of the formation of structurally well-defined hydrogen bonds that can provide specificity amongst ligands [22] and a significant change in the conformational entropy of the system. [23] An analysis of protein–protein interactions and the effect of amino acid mutations on their energetics has revealed that the observed differences in binding affinity are consistent with the changes in binding energy from the direct contact between each subunit of the complex and indirect changes due to the release of water molecules near the mutation site. [24] Failure to account for the effect of these water molecules can result in an underestimation of the calculated binding affinities.

Free energy simulations have been used to characterize the energetics of crystallographic water molecules. A threshold value of -50 kJ mol^{-1} has been found for distinguishing hydrated and empty sites of buried structural water molecules. [25] However, these buried water molecules can be considered to be an integral part of the protein structure and differ from water molecules mediating protein–ligand interactions in that they have a much slower exchange rate with the bulk solvent and are not accessible to an incoming ligand.

Such simulations have also been used to study the binding site of cytochrome P450cam. The protein complex with an inhibitor contains one crystallographically well-defined water molecule, which mediates and stabilizes the interaction by nearly -12 kJ mol^{-1}. However, the calculated free energy difference between the mono-hydrated and non-hydrated (as observed in the crystal structure) states of the complex with camphor (natural substrate) was found to favor the non-hydrated state by nearly -16 kJ mol^{-1}. [26] In another study, the hydration of the empty binding site of P450cam was calculated to be more energetically favorable with five to six water molecules than with the maximum possible of ten water molecules. [27] The authors later mutated the bound camphor to six water molecules, which hydrated the binding site fully, estimating the associated free energy of binding of camphor to be just over -29 kJ mol^{-1}. [28] These calculations reveal the crucial role that water molecules in the binding site of a protein can have in determining the energetics of ligand binding.

From the perspective of drug design, the concept of replacing and mimicking such crystallographically determined bound water molecules has become widespread. [29] The classic example in drug design is that of the active site of HIV protease, where replacement of a water molecule by a carbonyl group on a cyclic urea inhibitor contributed to an increase in the entropy by releasing the ordered bound water molecule. [30] However, the replacement of a bound water molecule by a chemical group on a ligand does not necessarily result in a decrease in the free energy of binding. [31]

Implicit hydration has explained the observed selectivity of ketoprofen and two structural analogues for two cyclooxygenase isozymes. [32] The water structure in the interior of the active sites is flexible and can easily accommodate changes in ligand structure as well as guide specificity. There are also cases where natural substrates [33] and designed inhibitors [34] have been shown to be able to include and/or conserve water-mediated contacts, instead of trivially replacing the water molecules. These observations suggest that a proper evaluation of the free energy changes involved in protein–water–ligand interactions may be needed in order to rationalize whether the replacement of a water molecule by an incoming ligand is advantageous energetically.

It becomes clear that the first step when deciding whether to consider crystallographically determined water molecules is to choose which water molecules are relevant. Attempts have been made to predict such ordered hydration sites by modular neural networks using protein sequence information. [35] A genetic algorithm has been reported to predict polar ligand interactions as well as those interactions mediated by conserved water molecules in proteins. [36] The temperature B-factors of water molecules, the number of protein–water hydrogen bonds and the density and hydrophilicity of neighboring protein atoms were used to discriminate between bound and displaced water molecules. This effect was found to be independent of the chemical nature of the ligand, while the protein microenvironment of each water molecule seemed to be the dominant influence.

A cluster analysis of consensus water sites in thrombin and trypsin has shown how these sites are conserved amongst serine proteases and how they contribute to ligand specificity. [5] It was found that highly conserved water micro-clusters generally had more neighboring protein atoms, were in a more hydrophilic environment, made more hydrogen bonds to the protein and were also less mobile. Water sites that could be identified as conserved in the thrombin structures were not identified as such in the trypsin structures, and vice versa, providing a list of water sites that might contribute to ligand discrimination. There were also significant overlaps between the thrombin and the trypsin conserved-water sites, likely to be associated with ligand selectivity.

Finally, empirical relationships between structural properties of the protein–solvent interface have been found by statistical analyses of crystal structure surveys. A correlation between occupancy and water temperature B-factors in protein crystal structures was found, [37] as well as a correlation between accessibility to internal cavities and ligand-binding sites with triads of atoms of comparable B-factors. [38] An estimate of the number of water molecules that can be expected in a protein crystal structure has been reported, [39] where a multivariate linear regression analysis was carried out to model the relative number of water molecules per number of protein atoms (roughly one water molecule per residue) in terms of several structural properties.

Water molecules bridging the interaction between a protein and a ligand can be included in drug design and molecular modeling strategies in different ways. The placing of explicit water molecules at favorable positions in the protein–ligand interface has been shown to guide and improve the docking of the ligand in FLEX-X, a program for protein–ligand docking with an incremental construction approach. [40] The water molecules included in the binding site interacted with fragments of the ligand being constructed if they were able to form additional hydrogen bonds with the ligand. Consequently, the steric constraints imposed by these water molecules as well as the geometry of the hydrogen bonds were used to optimize the ligand binding mode. This is an example of how water molecules may influence the geometry and free energy landscape of a ligand-binding site. At the same time, a strategy for the incorporation of water molecules inside a ligand-binding site into a three-dimensional quantitative structure–activity relationship (QSAR) analysis has been reported. [41] Such a procedure included crystallographic water molecules as part of the ligand structure and the results showed an improvement in the predictive ability of the models.

The knowledge of the effects that explicit water molecules can have on ligand binding and specificity should be considered in drug design strategies for a better evaluation of protein–water–ligand interactions, as well as the incorporation of new chemical features into the ligands being generated. Ligands might need lower numbers of polar or charged groups (hydrophilic contacts would be mediated through hydrogen bonds formed by the water molecules present), as well as smaller molecular weights as fewer hydrophobic contacts would be necessary. These effects arise from the observation that, when databases of dissimilar molecules are screened for complementarity to receptors of known structures, failure to consider ligand solvation often leads to putative ligands that are too highly charged or too large because of an overestimation of hydrophilic and hydrophobic protein–ligand interactions, respectively. [42]

Water molecules found in a protein crystal structure can be considered as part of the binding site of a protein where de novo assembly of a ligand will take place, the docking of a ligand will be carried out, or the magnitude of electrostatic, hydrogen-bonding and hydrophobic interactions between a ligand and a protein will be computed. Furthermore, water molecules can be considered to be mobile in the way they behave dynamically in their microenvironment, or fixed at certain locations where they appear to be conserved. The method described in this paper has been applied to selecting water molecules for de novo ligand design, where their presence modulated the chemical diversity of the designed ligands through the hydrogen-bonding and steric constraints they imposed. [43]

The present work deals with a strategy for incorporating crystallographically observed water molecules into molecular modeling methods, by considering their simple structural properties and determining their statistical significance. We have searched for a relationship between the structural properties of those water molecules observed in the same positions between different X-ray crystal structures of the same protein, in an attempt to predict the probability of a water molecule being bound to the protein surface at the same hydration site. Our aim has thus been to determine conserved water molecule positions that can be used to modify the shape and chemical properties of the binding site of a protein. This in turn can allow for a more realistic scoring of protein–ligand interactions, a more accurate determination of ligand binding modes and the modulation of chemical diversity in structure-based drug design.

Materials and methods

The proteins that were initially selected for both the calibration and testing sets (Tables 1 and 4, respectively) were those that have crystal structures for both their free and complex forms (with one or more ligands). Only crystal structures with a resolution better than 2.5 Å were considered. Furthermore, the proteins were chosen so as to minimize the effects on the hydration structure of any conformational changes in the binding site. Consequently, only those proteins that showed little or no geometric variation around their binding sites were considered. The crystal structures cover different levels of hydration: from fully hydrated binding sites (such as penicillopepsin, shown in Fig. 1 in Results and discussion) to binding sites with few water molecules (such as the lipid binding protein). The binding sites of the proteins chosen exhibit different shapes and sizes, as well as different types of bound ligands. There are several proteins that have small to medium-sized well-defined binding sites (such as cutinase, xylose isomerase, galactose/glucose binding protein, proteinase A, Rhizopuspepsin, cholesterol oxidase and dihydrofolate reductase), others that have large open binding sites (such as penicillopepsin, RNase A, thermitase and lipid binding protein) and others that have superficial ill-defined binding sites (such as actinidin, and the Fv fragment of mouse monoclonal antibody D1.3).

The selected protein data sets thus cover a range of typical examples of protein binding sites, to account for the different conditions that are likely to be found in protein–ligand crystal structures. Table 1 contains the names and other information for the 25 protein pairs analyzed, as well as counts for the different classes of water molecules. Some of the proteins analyzed had more than one complex with a number of ligands, and these were treated independently. For example, in the case of cutinase we analyzed a crystal structure of the free protein (1cex) and two crystal structures of complexes with two different ligands (1xzl and 1xzm). In these cases, the same atoms have been used to define the binding sites of the free enzyme and the several complexes. Water molecules excluded by one ligand in one of the complexes but not in another complex were not included in the analysis due to the statistical noise these would introduce (they would contribute with the same property values to both the bound waters category and also to the displaced water molecules category which we are trying to separate). Table 2 contains a data set for the evaluation of the performance of WaterScore. It has previously been used elsewhere by a related study for the testing of prediction of conservation of water molecules. [36]

The structural properties of water molecules in the free (uncomplexed) forms of the proteins which were analyzed were the temperature B-factors, the hydrogen-bonding energies of the water–protein and water–water contacts, the solvent-accessible (SASA) and solvent-contact surface-areas (SCSA) and the number of protein atomic contacts of each water molecule (NPAC). The temperature B-factors of the neighboring protein atoms and their corresponding SASA were also evaluated.

Table 1 Protein–water data calibration set for WaterScore

Protein pairs (free/complex PDB codes)	Protein name	Ligand name	Resolution (Å)	Bound waters	Displaced waters
1cex/(1xzl, 1xzm)	Cutinase	*N*-Hexylphosphonate-2-ethyl ester	1.0, 1.69, 1.75	2	1
1xyz/1xyb	Xylose isomerase	Xylose	1.81, 1.96	3	0
3app/(1ppk, 1ppm)	Penicillopepsin[a]	Statine derivative (1ppk), CBZ–Ala–Ala–Leu(P)–(O)Phe–OMe (1ppm)	1.8, 1.8, 1.7	4	1
1gcg/(2gbp, 3gbp)	Galactose/glucose binding protein	D-Glucose	1.9, 1.9, 2.4	3	0
2sga/(3sga, 4sga, 5sga)	Proteinase A (serine proteinase)	Phenyl alaninal, ACE–Pro–Ala–Pro–Phe, ACE–Pro–Ala–Pro–Tyr	1.5, 1.8, 1.8, 1.8	2	0
2apr/3apr	Rhizopuspepsin (aspartic proteinase)	Reduced phenyl alaninal	1.8, 1.8	2	0
2act/1aec	Actinidin	2 ([*N*-(L-3-*trans*-Carboxyoxirane-2-carbonyl)-L-leucil]-3 amido(4-guanido) butane)	1.7, 1.86	2	0
3dni/2dnj	DNase I	DNA fragment	2.0, 2.0	2	0
3cox/1coy	Cholesterol oxidase	Dehydroisoandrosterone	1.8, 1.8	9	1
1rbx/(1eow, 1rar, 1ras, 1rnc, 1rca, 1rbw)	Rnase A[b]	Uridylyl(2',5') guanosine (non-productive binding), acetylaminoethyl-naphtyl-amine sulfonate, citidylyl-2',5'-phosphoryl guanosine, deoxycytidylyl-3', 5'-guanosine, guanidinium.	1.69, 2.0, 1.9, 1.7, 1.5, 1.9, 1.69	2	1
1thm/2tec	Thermitase	Eglin-C	1.37, 1.98	0	1
1lib/1lic	Lipid binding protein	Hexadecanesulphonic acid	1.7, 1.6	0	1
1vfa/1vfb	Fv fragment of mouse monoclonal antibody D1.3	Hen egg lysozyme	1.8, 1.8	0	2
5dfr/(6dfr, 7dfr)	Dihydrofolate reductase	$NADP^+$, $NADP^+$+folate	2.3, 2.4, 2.5	0	1

[a] See Fig. 1
[b] See Fig. 2

Table 2 Testing set (cf. [36]) for WaterScore

Protein (free/complex PDB codes)	Ligand name	Resolution (Å)	Bound waters (correctly predicted/total)	Displaced waters (correctly predicted/total)	Waters removed by clashes (with protein/ with ligand)
Cyclodextrin glycosyl-transferase (1cgt/1cgu)	Glucose	2.0, 2.5	10/13	1/2	9 (4)
Trp repressor, DNA-binding regulatory protein (2wrp/1tro)	Operator	1.65, 1.9	3/4	3/3	9 (1)
Concanavalin-A (2ctv/5cna)	α-Methyl-D-mannopyranoside	1.95, 2.0	0/3	3/3	7 (3)
Dihydrofolate reductase (1dr2/1dr3)	$NADP^+$+biopterin	2.3, 2.3	9/9	2/9	0 (2)

Temperature B-factors (Bf)

The isotropic temperature B-factors provide an estimate of the atomic mobility within a crystal structure through the calculation of the mean displacement $\bar{U}$ of an atom by the relationship $B=8\pi^2\bar{U}^2$. Consequently, water molecules with high B-factors are less tightly bound to the protein surface as they are more mobile. The B-factors of water molecules and neighboring protein atoms were read directly from the PDB files. [44] The module PROCHECK in the program WHATIF [45] was used to investigate the quality of the crystal structure and to ensure that the B-factors were not unrealistically high or low.

Hydrogen-bond energies (WHBE)

The module HB2NET in the program WHATIF [45] was used to optimize the hydrogen atom positions of both the

protein and the water molecules in each crystal structure. The method scores hydrogen bonds using a special force field developed from a database of accurately determined small molecule structures. [46] Hydrogen bonds are given a score between 0 and 1, where 1 represents an energy value of 25 kJ mol^{-1} for an ideal (strong) hydrogen bond. The score for a particular hydrogen bond is determined from the donor/acceptor types, the hydrogen-bond donor–acceptor distance, the hydrogen-bond angle and the hydrogen-bond distance. The overall hydrogen-bond energy of the protein and water molecules is optimized iteratively by searching for the best possible position for all the hydrogen atoms simultaneously; this also takes into account the cooperativity affecting chains and networks of hydrogen bonds. During the optimization procedure, histidine, asparagine and glutamine side-chains are allowed to flip 180°, as crystallographic determinations cannot distinguish between the two rotamers. The total hydrogen-bond energy of a water molecule (WHBE) was then calculated as the sum of the energies of all water–protein and water–water hydrogen bonds. In the case of water–water hydrogen bonds, these were only considered when both water molecules involved were retained in the two crystal structures being compared.

Solvent-accessible surface-area (SASA)

The SASA for each water molecule was computed using the program NACCESS 2.1.1, [47] which calculates the atomic accessible surface defined by rolling a probe of a given size around the van der Waals surface and following the coordinates of the center of the probe. [48] This is a measure of the accessibility of a water molecule to the outer bulk aqueous environment. Less accessible water molecules are located in deeper crevices or grooves on the protein surface. The radius of the rolling probe used was 1.2 Å (cf. 1.4 Å is the radius of a water molecule), which has been used elsewhere for the exploration of protein surfaces [16] (cf. 1.25 Å [47]). The SCSA, in which the coordinates on the surface of the probe (rather than its center) are taken, was also measured. The SASA and SCSA of protein atoms in contact with a given water molecule were also computed.

Number of protein atomic contacts (NPAC)

A cutoff of 3.5 Å from the center of each water molecule was used to determine the number of atomic contacts with protein atoms. This is related to the local atomic density and to the local van der Waals interactions between a water molecule and the protein surface.

Since we were only interested in those water molecules residing in the binding sites of the proteins considered, we used a cutoff distance of 7.0 Å from any ligand atom to extract the binding sites. Water molecules within 3.5 Å of the protein atoms in the binding sites were also extracted. The module CHKWAT in WHATIF [45] was used to remove any water molecules that had coordinates too close to a non-water symmetry-related atom. The binding sites of the two proteins being considered were then superimposed by minimizing the atomic root mean square deviation (RMSD). Once the two binding sites and corresponding water molecules were in the same frame of reference, a cutoff of 0.5 Å was found to be the optimum for matching water molecules. If more than one water molecule was found to match the reference position, the one that was closest was then selected. Water molecules in the free form of the protein were then classified as *displaced* if they could not be matched with another water molecule in the complex form of the protein, and *bound* if they were matched successfully. Care was taken to avoid redundancy in the matching of water molecules, and a visual inspection was carried out in some cases to verify this.

An important difference between our approach and that of other authors [5, 36] is that water molecules that were expelled from the binding site due to steric interactions with a bound ligand were not included in our category of displaced water molecules. These water molecules are identified by superimposing the ligand of the complex form of a protein onto the free form of the same protein. Any water molecule that clashes with the ligand was therefore assumed to have been *sterically* displaced by the ligand upon binding. This procedure retained only water molecules in the complex form of the protein that could be classified as bound *after* ligand binding had taken place. Otherwise, there would have been an uncertainty as to which category the expelled water molecules belonged to, since it would not be possible to decide whether they would have been non-sterically displaced or not by a ligand of a different shape and/or size. Those water molecules displaced by steric influences (or clashes) with the ligand or protein cannot be included in the statistical analysis because they are not exhibiting their behavior as they would in normal circumstances, that is, without steric clashes.

Multivariate statistical analysis

A multivariate logistic statistical analysis of the structural properties of water molecules described above was carried using the program Matlab. [49] This enabled us to generate a correlation model to discriminate water molecules, some into a "displaceable" class (likely to be lost upon ligand binding, also called "non-conserved"), and others into "bound" class (likely to remain bound to the protein upon ligand binding, also called "conserved"). A logistic regression procedure was used because of its ability to provide an estimate of the probability of such discrimination in the form of the response or dependent variable. The regression statistics are principally the G_m value, which can be tested against a χ^2 distribution function for statistical significance; and the $R_L{}^2$ value, which determines how well the estimated response

variable is predicted by the correlation between the various independent variables. A more detailed description of the multivariate logistic regression method can be found in the Appendix.

Results and discussion

Water molecules that were successfully matched and classified as bound, were coded as 1, whilst all others were classified as displaced and coded as 0, for the purposes of the dependent (response) variable in the statistically-fitted models. The total number of water molecules considered was 40. There were a total of 30 bound water molecules (coded 1), with one outlier (this had a value that deviated more than three standard deviations from the mean), and with a fraction of 0.7692 of the total sample. There were a total of nine displaced water molecules (coded 0), with no outliers, and with a fraction of 0.2308 of the total sample.

Figure 1 shows the binding site of penicillopepsin (3app) with its crystallographically determined water molecules and a superimposed ligand (from the complexed structure 1ppk). It can be seen that the ligand occupies the positions where many of the water molecules in the free form of the protein lie (shown in cyan). We have considered such water molecules as having been sterically displaced by the ligand in the complexed form of the protein. Two water molecules were seen not to have any hydrogen bonds with the protein surface (shown in white), and were therefore excluded from our analysis, because they lack the protein–water properties that we considered in this study. Two bound water molecules (shown in blue, coded 1), appearing in both the free and the complexed forms of the protein, bridge the interaction between the ligand and the protein. A single displaced water molecule (shown in yellow, coded 0) was observed in the free form of the protein, but could not be matched in the complexed form even though there are no steric clashes with the ligand.

Figure 2 shows another example taken from the protein data set that we analyzed. In this case, the binding site of RNase A (1rbx) with its crystallographically determined water molecules and several superimposed ligands are shown. As before, the ligand occupies the positions of some of the water molecules in the free form of the protein, while others are retained (bound) and one is displaced (not by steric clashes with any of the ligands).

The above two proteins exemplify the fact that we observed no obvious visual pattern for the location of either bound or displaceable water molecules. It was not necessarily true that bound water molecules were "sandwiched" between the ligand and the protein while displaceable water molecules were located on the surface of the binding site. It becomes clear that a quantitative analysis of a number of structural properties of water molecules is necessary to model their behavior upon ligand binding.

Table 3 shows the means and standard deviations for all the properties of bound and displaced water molecules

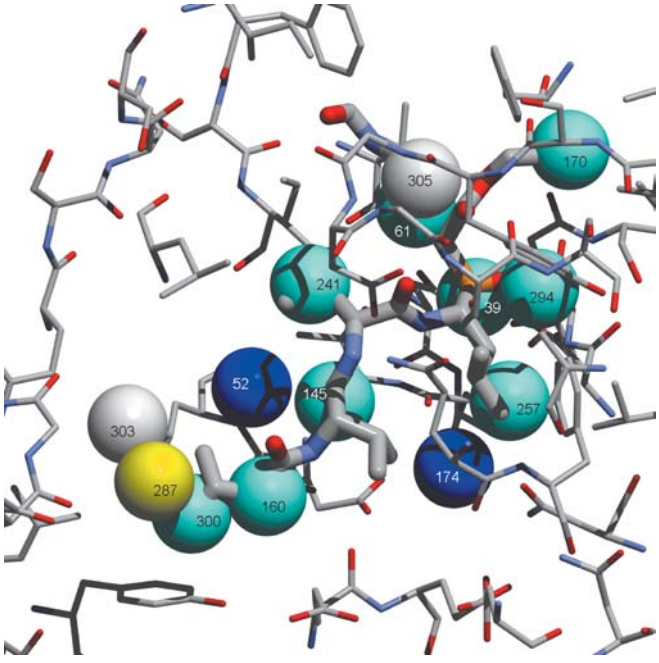

Fig. 1 The binding site (thin sticks) of penicillopepsin (3app) with its crystallographically determined water molecules (*spheres*) and superimposed ligand (in *thick sticks*, from complexed structure 1ppk). Water molecules sterically displaced by the ligand upon complexation are shown in *cyan*. Bound water molecules are shown in *blue*. Displaced water molecules are shown in *yellow*. Water molecules removed from the analysis due to a lack of hydrogen bonds to the protein are shown in *white*. WaterScore correctly predicted waters in *blue* as Probability=1 to remain bound and waters in *yellow* as Probability<1×10^{-20} to remain bound

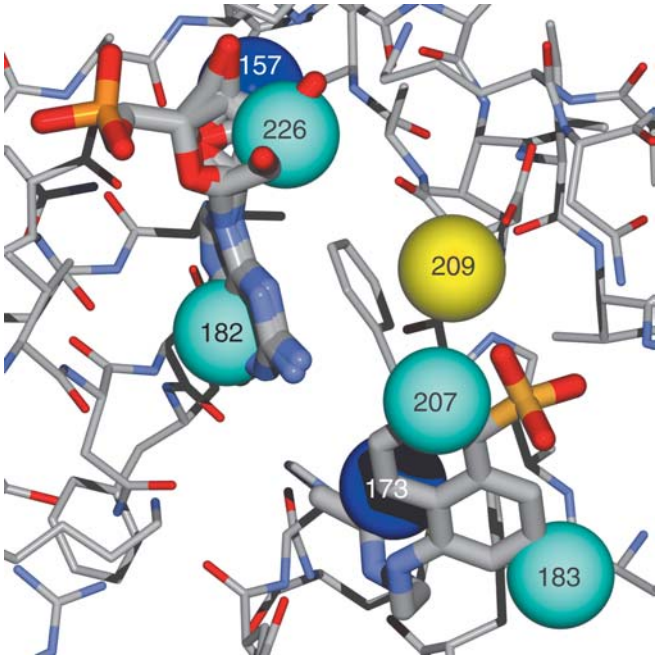

Fig. 2 The binding site of RNase A (1rbx) with its crystallographically determined water molecules and several superimposed ligands (from complexed structures 1eow, 1ras, 1rar, 1rnc, 1rca and 1rbw). Water molecules sterically displaced by the ligand upon complexation are shown in cyan. Bound water molecules are shown in *blue*. Displaced water molecules are shown in *yellow*

Table 3 Means and standard deviations for the structural properties of water molecules

Water molecules	B-factor	SCSA (Å^2)	NPAC	WHBE (kJ mol^{-1})
Bound	20.3±11.9	3.03±3.09	3.84±1.98	−38.66±16.55
Displaced	38.0±10.4	8.41±3.66	0.68±0.71	−29.22±17.60

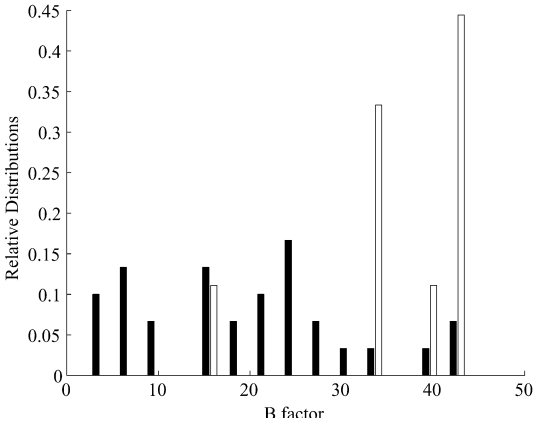

Fig. 3 Histogram of the distributions of values of the B-factors of bound (*solid bars*) and displaced (*open bars*) water molecules

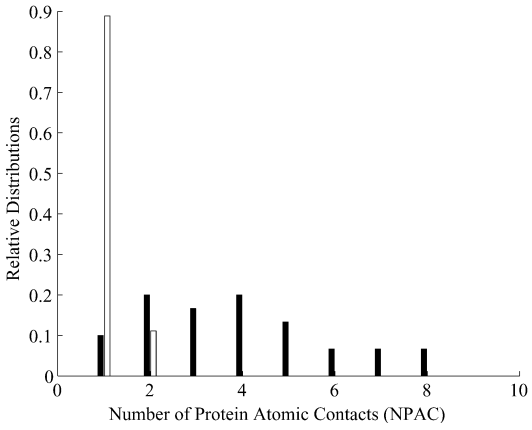

Fig. 5 Histogram of the distributions of values of the NPAC of bound (*solid bars*) and displaced (*open bars*) water molecules

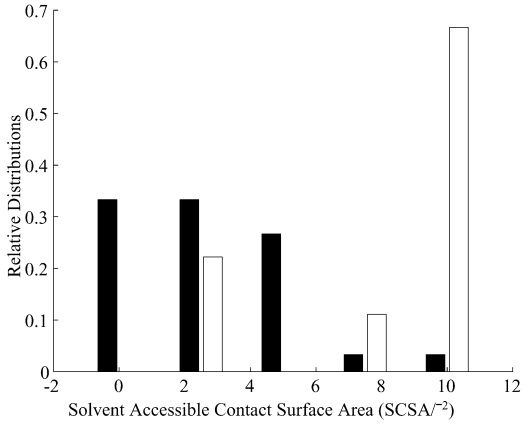

Fig. 4 Histogram of the distributions of values of the SCSA of bound (*solid bars*) and displaced (*open bars*) water molecules

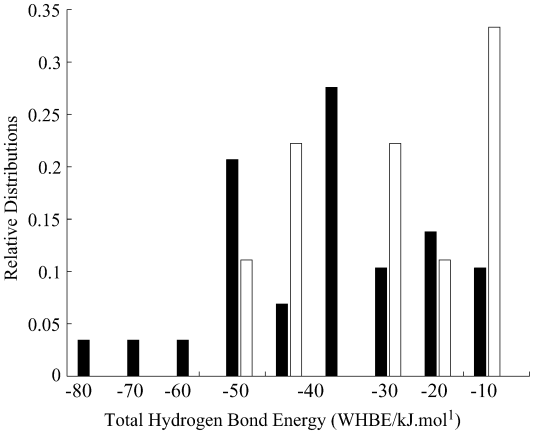

Fig. 6 Histogram of the distributions of values of the WHBE of bound (*solid bars*) and displaced (*open bars*) water molecules

considered. We can see that the means for the B-factor, SCSA and WHBE are higher for displaced than for bound water molecules, while NPAC is greater for bound water molecules. As a consequence, bound water molecules tend to have low B-factors, small surface areas exposed to the solvent, a large number of atomic contacts with protein atoms and low hydrogen bond energies.

Histograms of the distributions of values for the above structural properties for all bound and displaced water molecules are shown in Figs. 3, 4, 5 and 6. From these figures we can see that bound water molecules have distributions of B-factors, SCSA and WHBE that are shifted towards lower values as compared to those of displaced water molecules; the distributions of NPAC values clearly show the opposite behavior. All these results suggest that an important condition for water molecules to remain bound to the protein surface upon ligand binding is to be buried deep in a crevice or groove in the binding site while surrounded by many protein atoms and making many hydrogen bonds, which also restricts their mobility and the surface they expose to the solvent.

We calculated correlation factors between all four variables: B-factor/SCSA (0.605), B-factor/NPAC (–0.501), B-factor/WHBE (–0.230), SCSA/NPAC (–0.349), SCSA/WHBE (–0.432) and NPAC/WHBE (0.086). We can see that the level of correlation between any pair of variables is sufficiently low for their use in a multi-parametric statistical analysis. These variables were consequently fed into a logistic regression analysis, producing the set of statistics shown in Table 4.

Table 4 Multivariate logistic regression statistics. Model 1 is the thee-variable model (*A* in Eq. 2) and Model 2 is the four-variable model (*A*2 in Eq. 4)

Model	D_m	D_0	G_m	R_L^2	$\chi^2_{95\%}$	$\chi^2_{70\%}$	$G_m/\chi^2_{70\%}$
1	1.35×10^{-6}	42.1359	42.1359	1.0	53.0991	42.1385	0.9999
2	2.29×10^{-6}	42.1359	42.1359	1.0	53.0991	42.1385	0.9999

The best logistic regression model was obtained with just three variables (B-factors, SCSA and NPAC). This model shows a strong correlation (R_L^2=1.0), and statistically significant evidence against the null hypothesis (of no correlation) at a 70% confidence level. The final equation obtained was

$$P(Y=1) = \exp[A]/(1+\exp[A]) \quad (1)$$

with

$$A = a - b_1\mathrm{Bf} - b_2\mathrm{SCSA} + b_3\mathrm{NPAC} \quad (2)$$

Here $P(Y=1)$ is the probability of a water molecule being classified as bound, and the coefficient values are: a=76.442, b_1=5.278, b_2=2.166, b_3=84.458.

We can observe in the previous equations that the B-factor and SCSA have an expected negative logistic relationship: the higher the values of B-factor and SCSA, the lower the probability of a water molecule being classified as bound. On the other hand, NPAC shows a positive logistic relationship: the lower the values of NPAC, the lower the probability of classifying the water molecule as bound.

Individual logistic regressions were also carried out individually for each of the variables in order to assess which of these had a larger weight in determining the overall correlation between each independent variable and *P* (probability). Although we observed that NPAC had the largest coefficient in the multivariate model, all three variables were required to produce a satisfactory model.

Our second best model included all four variables, namely those of the previous model and the total hydrogen-bond energy (WHBE) of each individual water molecule. From Table 3 and Fig. 6 we can see that bound water molecules have more negative WHBE than displaced water molecules, reflecting increased hydrogen bonding with the protein surface and/or other water molecules. This model also showed statistical significance at a 70% confidence level, constituting an alternative to the previous model. Its final equation was

$$A_2 = c - d_1\mathrm{Bf} - d_2\mathrm{SCSA} + d_3\mathrm{NPAC} - d_4\mathrm{WHBE} \quad (3)$$

with c=44.683, d_1=4.165, d_2=4.017, d_3=54.439, d_4=0.998 and where A_2 can be substituted for A in Eq. (1) above.

We can see that Bf, SCSA and NPAC have the same logistic relationships seen in Eq. (2): negative for Bf and SCSA, and positive for NPAC. However, their weights in this logistic regression model are somewhat different (i.e., about twice as large for SCSA in A_2 than in A, and roughly half as large for NPAC in the same comparison). This can be due to the fact that Eq. (3) now incorporates WHBE. The negative logistic behavior of WHBE confirms that the lower the hydrogen bond energy, the higher the probability of a water molecule being classified as bound. Eq. (2) might be preferable to Eq. (3) since it has fewer variables; however, Eq. (3) can provide a smoother model, as the magnitude of the constant coefficient and the weight of NPAC (the only discrete variable considered) are smaller in relation to the other terms.

A principal components analysis (PCA) was carried out to identify the variables that contributed the most to the variance of the data points. Figure 7a shows the plot of the objects (water molecule observations) on the axes of the two main principal components.

Figure 7b shows the loading plot of the variables and how they contribute to the principal components (PC). B-factor and WHBE (with also a negative contribution from SCSA) are those standing out in PC1 and PC2, respectively.

Clearly, the two distinct groups of water molecules we aim to separate are distinguished by the PCA study. With the exception of point 38, conserved water molecules have values higher in principal component PC2 (mainly hydrogen bond energy and negative solvent contact surface area), while those displaced have values higher in principal component 1, PC1 (mainly B-factor). PC1 explains 82.77% of the variance of the data, while PC2 explains 8.53% (together explaining 91.3%). PC3 accounts for 4.36% (the first three principal components explaining 95.7% of the variance in the data), and PC4 for another 4.31% (the total sum of four principal components is 99.97%).

From Fig. 7b it is seen that at least three variables need to be taken into account for an acceptable description of the data. This test provides strength of argument to the distinction between these two classes of water molecules by three variables.

A logistic regression using only any two variables did not produce results of the statistical significance of Eq. (1) using either Eq. (2) or Eq. (3). Therefore, we chose to use Eq. (2) (three-variable) as our model for testing and for implementation in WaterScore.

The water molecules in the binding site of the enzymes in the calibration (or training) set were scored with Eq. (2), and all bound or conserved water molecules scored a probability of 1.0 or very near to this value, while all displaced water molecules scored a probability of less than 1×10^{-20}. That is, they all are predicted correctly according to the calibration (see Fig. 1).

WaterScore was then tested on another set of enzymes for evaluation of its prediction scores. The results are presented in Table 2.

Overall, the results are encouraging, since the program performs reasonably well on new systems, though strong consideration needs to be taken to achieve a good superposition of binding sites. WaterScore scores water

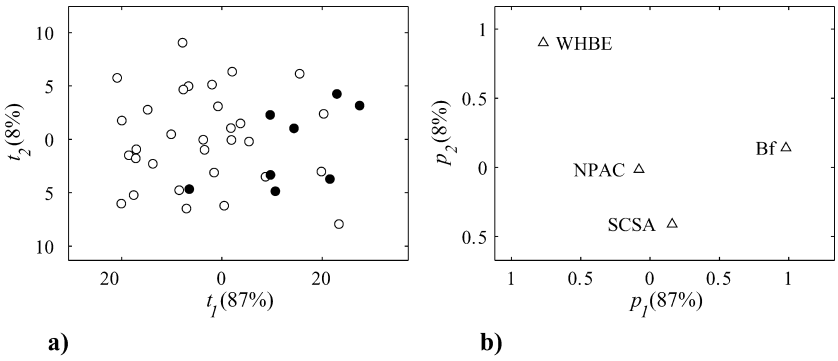

Fig. 7 **a** Plot of the water molecule observations along their two principal components in variance (from PCA). Open circles indicate the water molecules conserved and solid circles show those displaced. **b** Loadings plot of the variables included in the PCA of the water molecule observations

molecules from a range of probability values of 5.9×10^{-72} to 1.0 on a total of 46 water molecules tested for prediction. Considering a probability threshold of 1×10^{-20} to distinguish between conserved and displaced waters, and a threshold of 1.5 Å for the distance between water molecules, gives a prediction efficiency of 67.4%. This value is acceptable considering the widespread applicability of the program and method. If a looser (wider) threshold of 2.0 Å is allowed for the distance between water molecules, the efficiency improves to 71.7%.

Conclusions

We have obtained novel multivariate logistic models to establish a quantitative relationship between simple micro-environmental structural properties of water molecules in the empty binding site of a protein and the probability of observing the same (bound) water molecules after ligand binding. Our models make use of the B-factor, the solvent-contact surface area, the total hydrogen bond energy and the number of protein atomic contacts that a water molecule has, showing that bound water molecules are likely to have low B-factors, small SCSA, low WHBE and large NPAC. This is indicative of water molecules of low mobility, buried deep in crevices or grooves in the binding site of a protein. This provides a consistent approach to the inclusion of water molecules in protein binding sites across different biomolecular applications.

There are two advantages to having simple models for determining whether a water molecule will be displaced or remain bound upon ligand binding. The first one is that they are very fast methods that can easily be updated as new and better-resolved protein crystal structures become available. The second one is that such models allow for the immediate analysis of protein crystal structures for the judicious selection of water molecules to be included in protein–ligand docking and/or structure-based drug design. This should lead to the prediction of more accurate binding modes and free energies of binding as well as the modulation of chemical diversity in designed ligands.

We are currently extending our applications of this method to de novo drug design [45] and ligand docking for a number of protein targets where water molecules are likely to play an important role in ligand-binding specificity and plasticity.

Lists of selected scorings and typical output produced by WaterScore can be seen at http://www.cus.cam.ac.uk/~atg21

Acknowledgements ATGS would like to thank Consejo Nacional de Ciencia y Tecnología (CONACyT, México) for the award of a postgraduate scholarship and the CVCP of the Universities of the UK for an Overseas Research Scheme award. RLM is also a Research Fellow of Hughes Hall, Cambridge. We also thank Mr. Benjamin Carrington for his valuable help in the production of some of the figures, Dr. Per Kållblad for help and discussion on PC analysis, and Miss Eva-Liina Asu for proof-reading a draft of the manuscript.

Appendix

We provide a brief outline of multivariate logistic regression analysis. [50, 51, 52] For a binary dependent variable Y that can take values of either 0 or 1, its mean is the proportion of cases of the higher value (1), and the predicted value of the dependent variable (the conditional mean, given the value of the independent variable X and the assumption that Y and X are linearly related) can be interpreted as the *predicted probability* that an observation falls into such higher value. By definition, the predicted probability lies between 0 and 1. The general shape of the relationship between the probability $P(Y=1)$ and the independent variable X is that of an "S curve", as depicted in Fig. 8.

Instead of predicting the arbitrary value associated with the dependent variable Y, it may be useful to predict the probability that a given observation (as defined by a set of independent variables) will be classified into one of the two values of the dependent variable. Naturally, if we know $P(Y=1)$, we immediately also know the probability of $P(Y=0)$ as $P(Y=0)=1-P(Y=1)$.

If the probability that $Y=1$ is modeled as $P(Y=1)=\alpha+\beta X$, its predicted values may be less than 0 or greater than 1. The first step to avoid this is to replace the probability that $Y=1$ with the *odds* that $Y=1$. The odds that $Y=1$, written Odds($Y=1$), is the ratio of the probability that $Y=1$ to the probability that $Y\neq1$. Odds($Y=1$) is then equal to $P(Y=1)/[1-P(Y=1)]$. Unlike $P(Y=1)$, the odds has

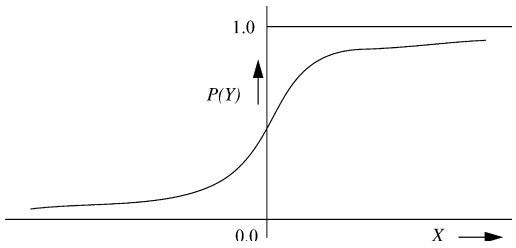

Fig. 8 The logistic curve model for the probability $P(Y=1)$ of a binary dependent variable showing a positive correlation with the independent variable X

no fixed maximum value, but like the probability, it has a minimum value of 0.

One further transformation of the odds produces a variable that varies, in principle, from negative infinity to positive infinity. The natural logarithm of the odds, $\ln\{P(Y=1)/[1-P(Y=1)]\}$, is called the *logit* of Y, and is written logit(Y). This function becomes negative and increasingly large as the odds decrease from 1 to 0, and becomes positive and increasingly large as the odds increase from 1 to infinity. By using the natural logarithm of the odds that $Y=1$ as the dependent variable, one no longer has the problem that the estimated probability may exceed the maximum or minimum possible values for the probability (see Fig. 8). The equation for the relationship between the dependent variable and a number of independent variables can be then expressed as

$$\text{logit}(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{4}$$

Calculating back the odds as $\text{Odds}(Y=1)=\exp[\text{logit}(Y)]$ gives us

$$\begin{aligned}\text{Odds}(Y=1) &= \exp\{\ln[\text{Odds}(Y=1)]\} \\ &= \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)\end{aligned} \tag{5}$$

A change of unit in X_i multiplies the odds by $\exp(\beta)$. The odds can be converted back to the probability that $Y=1$ by the formula $P(Y=1)=\text{Odds}(Y=1)/[1+\text{Odds}(Y=1)]$, producing the equation

$$P(Y=1) = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)} \tag{6}$$

For any given case, $\text{logit}(Y)=\pm\infty$. This ensures that the probabilities estimated will not be less than 0 or greater than 1. Because the linear form of the model (Eq. 4) can have infinitely large or small values for the dependent variable, ordinary least squares (OLS) cannot be used to estimate the parameters β_i. Instead, maximum likelihood techniques are used to maximize the value of the log likelihood (LL) function, which indicates how likely it is to obtain the observed values of Y, given the values of the independent variables and the parameters α, β_1, ..., β_k. Unlike OLS, which is able to solve directly for the parameters, the solution of the logistic regression model is found by iterating the estimation until the solution converges when the change in the likelihood function is negligible (for the present study, we used a threshold of 1×10^{-6}, in the routine logitfit.m [53] for Matlab [49]).

Twice the negative of LL has approximately a χ^2 distribution, which allows one to test the goodness of fit of a model. The value of −2LL for the logistic regression model with only the intercept included is designated D_0 to indicate that it is the −2 log likelihood statistic with none of the independent variables in the equation. It is analogous to the sum of squares (SST), in linear regression analysis. D_m is analogous to the error sum of squares (SSE) in linear regression analysis, and is sometimes called "deviance", and is twice the negative LL function with the intercept as well as all the independent variables included. D_m is used as an indicator of how poorly the model fits all of the independent variables in the equation. D_m is analogous to the statistical significance of the unexplained variance in a regression model. The most direct analogue in logistic regression analysis to the regression sum of squares (SSR) in linear regression analysis is the difference between D_0 and D_m:

$$G_m = \chi^2 = (D_0 - D_m) \tag{7}$$

G_m is analogous to the multivariate F-test for linear regression, as well as the regression sum of squares. Treated as a χ^2 statistic, G_m provides a test of the null hypothesis that $\beta_1=\beta_2=...=\beta_k=0$ for the logistic regression model. If G_m is statistically significant (with, for example, $p<0.05$, a 95% confidence level), then the null hypothesis (of random correlation) is rejected and one can conclude that the model allows us to make predictions of $P(Y=1)$.

A natural choice for comparing the strength of the relationship between variables is the analogy to R^2 as the sum of the squares of the residuals over the total sum of squares (SST), SST=SSR / SST, in a linear regression model. $R_L{}^2$ is a proportional reduction in χ^2 or a proportional reduction in the absolute value of the LL measure.

$$R_L^2 = G_m/D_0 \tag{8}$$

This statistic indicates by how much the inclusion of the independent variables in the model increases the goodness of fit D_0 to the χ^2 statistic. $R_L{}^2$ varies between 0 (for a model in which $G_m=0$, $D_m=D_0$ and the independent variables are useless in predicting the dependent variable) and 1 (for a model in which $G_m=-2LL$ and $D_m=0$ and the model predicts the dependent variable with perfect accuracy).

References

1. Giacovazzo C, Monaco HL, Viterbo D, Scordari F, Gilli G, Zanotti G, Catti M (1992) Fundamentals of crystallography. Oxford University Press, Oxford, pp 583–584
2. Jeffrey GA (1994) J Mol Struct 322:21–25
3. Purkiss A, Skoulakis S, Goodfellow JM (2001) Philos Trans R Soc London Ser A 359:1515–1527

4. Chung E, Henriques D, Renzoni D, Zvelebil M, Bradshaw JM, Waksman G, Robinson CV, Ladbury JE (1998) Struct Folding Design 6:1141–1151
5. Sanschagrin PC, Kuhn LA (1998) Protein Sci 7:2054–2064
6. Lemieux RU (1996) Acc Chem Res 29:373–380
7. Nakasako M (1999) J Mol Biol 289:547–564
8. Faerman CH, Karplus PA (1995) PROTEINS 23:1–11
9. Schwabe JWR (1997) Curr Opin Struct Biol 7:126–134
10. Carrell HL, Glusker JP, Burger V, Manfre F, Tritsch D, Biellmann J-F (1989) Proc Natl Acad Sci USA 86:4440–4444
11. Baker EL, Hubbard RE (1984) Prog Biophys Molec Biol 44:97–179
12. Loris R, Langhorst U, De Vos S, Decanniere K, Bouckaert J, Maes D, Transhue TR, Steyaert J (1999) PROTEINS 36:117–134
13. Loris R, Stas PP, Wyns L (1994) J Biol Chem 269:26722–26733
14. Poornima CS, Dean PM (1995) J Comput-Aided Mol Des 9:521–531
15. Poornima CS, Dean PM (1995) J Comput-Aided Mol Des 9:500–512
16. Poornima CS, Dean PM (1995) J Comput-Aided Mol Des 9:513–520
17. Feig M, Pettitt BM (1998) Structure 6:1351–1354
18. Zhang X-J, Matthews BW (1994) Protein Sci 3:1031–1039
19. Mattos C (2002) Trends Biochem Sci 27:203–208
20. Esposito L, Vitagliano L, Sica F, Sorrentino G, Zagari A, Mazzarella L (2000) J Mol Biol 297:713–732
21. Teeter MM (1991) Annu Rev Biophys Chem 20:577–600
22. Swaminathan CP, Nandi A, Visweswariah SS, Surolia A (1999) J Biol Chem 274:31272–31278
23. Bhat TN, Bentley GA, Boulot G, Greene MI, Tello D, Dall'Acqua W, Souchon H, Schwarz FP, Mariuzza RA, Poljal RJ (1994) Proc Natl Acad Sci USA 91:1089–1093
24. Covell DG, Wallqvist A (1997) J Mol Biol 269:281–297
25. Zhang L, Hermans J (1996) PROTEINS 24:433–438
26. Helms V, Wade RC (1995) Biophys J 69:810–824
27. Helms V, Wade RC (1998) PROTEINS 32:381–396
28. Helms V, Wade RC (1998) J Am Chem Soc 120:2710–2713
29. Marrone TJ, Briggs JM, McCammon JA (1997) Annu Rev Pharmacol Toxicol 37:71–90
30. Lam PYS, Jadhav PK, Eyermann CJ, Hodge CN, Ru Y, Bacheler LT, Meek JL, Otto MJ, Rayner MM, Wong YN, Chang CH, Weber PC, Jackson DA, Sharpe, TR, Ericksonviitanen S (1994) Science 263:380–384
31. Mikol V, Papageorgiou C, Borer X (1995) J Med Chem 38:3361–3367
32. Palomer A, Pérez JJ, Navea S, Llorens O, Pascual J, García Ll, Mauleón D (2000) J Med Chem 43:2280–2284
33. Cherbavaz DB, Lee ME, Stroud RM, Koschl DE (2000) J Mol Biol 295:377–385
34. Finley JB, Atigadda VR, Duarte F, Zhao JJ, Brouillette WJ, Air GM, Luo M (1999) J Mol Biol 293:1107–1119
35. Ehrlich L, Reckzo M, Wade RC (1998) Protein Eng 11:11–19
36. Raymer ML, Sanschagrin PC, Punch WF, Venkataram S, Goodman ED, Kuhn L (1997) J Mol Biol 265:445–464
37. Carugo O (1999) Protein Eng 12:1021–1024
38. Carugo O, Argos P (1998) PROTEINS 31:201–213
39. Carugo O, Bordo D (1999) Acta Crystallogr Sect D 55:479–483
40. Rarey M, Kramer B, Lengauer T (1999) PROTEINS 34:17–28
41. Pastor M, Cruciani G, Watson KA (1997) J Med Chem 40:4089–4102
42. Shoichet BK, Leach AR, Kuntz ID (1999) PROTEINS 34:4–16
43. Mancera RL (2002) J Comp-Aided Mol Des 16:479–499
44. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
45. Vriend G (1990) J Mol Graph 8:52–56
46. Hooft RWW, Sander C, Vriend G (1996) PROTEINS 26:363–376
47. Hubbard SJ, Argos P (1995) Protein Eng 8:1011–1015
48. Lee B, Richards FM (1971) J Mol Biol 55:379–400
49. Matlab 5.0 (1999) The Math Works,
50. Menard SM (1995) Applied logistic regression analysis in series. In: Lewis-Beck MS (ed) Quantitative applications in the social sciences. Sage, Thousand Oaks, Calif.
51. Agresti A (1996) An introduction to categorical data analysis, Wiley series in probability and statistics, applied probability and statistics. Wiley, New York
52. Rice JA (1995) Mathematical statistics and data analysis, 2nd edn. Duxbury Press, Belmont, Calif.
53. Holtsberg A (1994) http://www.mathtools.net