*Journal of Computer-Aided Molecular Design* **18:** 89–100, 2004.



# The effect of tightly bound water molecules on the structural interpretation of ligand-derived pharmacophore models

David G. Lloyd[1,2,3], Alfonso T. García-Sosa[2], Ian L. Alberts[1], Nikolay P. Todorov[1] & Ricardo L. Mancera[1,*]

[1]*De Novo Pharmaceuticals, Compass House, Vision Park, Chivers Way, Histon, Cambridge CB4 9ZR, UK;* [2]*Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1PD, UK;* [3]*Present address: Molecular Design Group, Department of Biochemistry, Trinity College Dublin, Dublin 2, Ireland*

Received 1 December 2003; accepted in revised form 16 February 2004

*Key words:* binding site model, drug design, pharmacophore, water molecules

**Summary**

The importance of the consideration of water molecules in the structural interpretation of ligand-derived pharmacophore models is explored. We compare and combine results from recently introduced methods for bound-water molecule identification in protein binding sites and ligand-superposition-based pharmacophore derivation, for the interpretation of ligand-derived pharmacophore models. In the analysis of thymidine kinase (HSV-1) and poly (ADP-ribose) polymerase (PARP), the concurrent application of both methods leads to an agreement in the prediction of tightly bound water molecules as key pharmacophoric points in the binding site of these proteins. This agreement has implications for approaching binding site analysis and consensus drug design, as it highlights how pharmacophore-based models of binding sites can include interaction features not only with protein groups but also with bound water molecules.

## Introduction

Water molecules found in the crystal structure of a protein are either buried in its interior or bound to its surface. The former are thought to contribute to the conformational stability of the protein [1], while the latter are usually found in large numbers and are thought to contribute to the binding of ligands. It is important to realise that the identification and location of water molecules in the crystal structure of a protein can be problematic and that some of the water molecules found on the surface of a protein can be artefacts of the determination, particularly when they do not form a single hydrogen bond to any other atom [2].

Despite their possible importance in drug design and ligand–protein docking, ordered water molecules that appear in the crystal structure of the target protein are usually removed. An analogous situation is observed in the case of ligand-based drug design approaches, where a pharmacophore model is derived from a set of biologically active ligands but are water molecules rarely considered as potential matching interaction sitepoints in the binding site of the receptor protein. The reality is, however, that some of those water molecules seen in crystal structures effectively modify the shape of the binding site and interact through their hydrogen bonding groups, 'bridging' the interaction between the protein and a ligand and forming a complex hydrogen bonding network that stabilizes the protein–ligand interaction [3–7]. Consequently, the docking or the design of (new) ligands should take advantage of these water molecules so that compounds can bind effectively through both ligand–protein interactions and ligand–water interactions.

Displacing, mimicking and/or using bound water molecules is becoming increasingly widespread in drug design [8]. The replacement of a water molecule in the binding site of HIV protease by a carbonyl group

*To whom correspondence should be addressed. E-mail: Ricardo.Mancera@denovopharma.com

in a cyclic urea inhibitor contributed to a favorable increase in entropy by releasing this ordered tightly bound water molecule [9]. The displacement of a water molecule from the active site increased the potency of inhibitors of scytalone dehydratase [10]. However, replacing a tightly bound water molecule by a chemical group in a ligand does not necessarily decrease the free energy of binding [11], nor does the appearance of a new tightly bound water molecule bridging a ligand–protein interaction necessarily increase the binding affinity of the ligand [12]. Natural substrates [13] and designed inhibitors [5] have been shown to include and/or conserve water-mediated contacts, rather than replace the water molecules. It would thus seem that, in some cases, trapping water molecules within a protein–ligand complex can outweigh the unfavorable entropic cost of their restricted motion by providing favorable hydrogen-bonding interactions with both the protein and the ligand.

It is important for any drug design strategy or docking simulation to consider the effect of bound water molecules on the ligand binding mode and its associated ligand–protein binding energy. By placing explicit water molecules at favorable positions in the binding site of a protein, the success of a docking algorithm was significantly improved through the additional hydrogen bonds introduced between the water molecules and the ligand [14]. Tightly bound water molecules have also been used to improve the performance of the virtual screening of a large data set of organic compounds [15,16] and to distinguish the binding of pyrimidines and purines to herpes simplex virus thymidine kinase [16]. Water molecules within a binding site have also been included into a three-dimensional quantitative structure–activity relationship (QSAR) analysis, improving the predictive ability of the models [17]. Bound water molecules have been used with the de novo design of ligands, exhibiting a marked influence on the chemical diversity and binding modes of the generated ligands [18].

It is known that water molecules can occupy the same positions in crystal structures of the same protein obtained under different conditions [19] and/or with different ligands [20–22], or in a set of structurally related proteins [23–27]. The prediction of such ordered hydration sites has been attempted by using neural networks on protein sequence information [28]. Genetic algorithms have also been used to predict interactions mediated by conserved water molecules as well as polar ligand interactions [29]. A cluster analysis of consensus water sites in thrombin and trypsin revealed both the common and differential conservation of these sites between serine proteases and their role in ligand selectivity and specificity [30].

The first step before considering the use of tightly bound water molecules is deciding which water molecules found in a crystal structure are relevant. A recently developed multivariate logistic regression method called WaterScore can readily estimate the probability of observing a water molecule in the same location in the crystal structure of the ligand-complexed form of a protein [31]. Structural properties such as 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 effectively between bound and displaceable water molecules.

Despite the evidence that has been accumulated over the years from the above structural approaches, there is still little mention in the literature of the role that water molecules can potentially play in the interpretation or enhancement of a pharmacophore derived from a set of known active ligands [32–34]. We are aware that no ligand-based approach, on its own, guarantees that the conformations of the ligands superimposed by the method are identical to their bio-active conformations – thus we have adopted an integrated ligand and structure-based treatment in this work, seeking to address this issue by comparing a recently introduced structure-based method for predicting tightly bound water molecules [31] with another newly developed method for deriving a binding site model based on the superposition of a set of ligands [35]. Our aim is to provide specific ideal system examples which demonstrate where the prediction of tightly bound water molecules matches interaction points in a pharmacophore-derived binding site model that would otherwise have been assumed to be protein hydrogen bonding groups. This should open the door for water molecules to be properly considered as potential interaction features in a pharmacophore model.

## Materials and methods

Our work is divided into two parts. In the first part we use a set of ligands for each of two target proteins – poly ADP ribose polymerase (PARP) and herpes simplex virus thymidine kinase (HSV-1) – to derive a pharmacophore-based binding site model, which includes a number of hydrogen-bonding interaction fea-

tures. Some of these features have no immediate apparent matching protein sitepoints when considered in the corresponding protein binding site. In the second part we make a prediction of tightly bound water molecules within the binding site of the proteins and produce an overlay to identify predicted bound water molecules. The method gives us information about the degree of conservation of these bound-water molecule positions within the protein and if they are sterically displaced by any (or all) of the ligands or whether they mediate the interaction between the protein and the ligands. Interpreting the pharmacophore model with due consideration of the predicted bound-water molecule positions facilitates a successful match for the 'orphan' pharmacophore features and re-emphasises the importance of the matched water molecules in both ligand binding and pharmacophore interpretation.

*Ligand-based binding site model derivation*

To prepare a binding site model from known ligands we have applied a recently introduced method called Quasi2 that can produce ligand-based pharmacophore-derived binding site models for use in virtual screening through molecular superpositioning [35].

This ligand superposition method, Quasi2, produces ligand-derived site models through the optimisation of molecular similarity within a set of ligands with respect to those features known to be important in binding to biomolecular targets as a function of ligand conformation, ionisation state and tautomeric state. The algorithm is particularly suited to identifying partial similarities in a series of structures, thus reflecting the situation in which different ligands exploit different interactions with the binding site. This ability to match and score ligands based on their (partial) three-dimensional similarity has the potential to be exploited as a virtual high-throughput screening tool, where it can be run quickly enough to process large numbers of compounds, and still accurately enough to discern active compounds from inactive compounds. Quasi2 overlays multiple flexible molecules so that corresponding features on different molecules are identified and effectively superposed. The strategy comprises the assignment of molecular descriptors, ligand refitting, superposition scoring and optimisation.

Molecular descriptors corresponding to ligand binding features are assigned to the non-hydrogen atoms in each molecule. In this work, four descriptors are defined corresponding to hydrogen bond donor, acceptor, steric and aromatic properties. Each hydrogen-bond donor point is positioned at a distance of 3.0 Å from the corresponding donor atom along the donor-H vector. Hydrogen-bond acceptor points coincide with the corresponding hydrogen-bond acceptor atoms. These assignments are based on the results of crystal surveys of hydrogen-bonding interactions [36], which demonstrate pronounced directional character from a distribution of donor atoms towards a more positionally invariant acceptor. Steric and aromatic points are positioned to coincide with the corresponding steric and aromatic atoms. Initially, random conformations are generated for the ligands and their centres are made coincident. One of the ligands is chosen as an anchor, and all other molecules are fitted onto the selected structure and the alignment is scored. The initial superposition is likely to be far from optimal and, thus, the alignment is modified and the ligands refitted. The new overlay is accepted with a certain probability based on its score. These steps are repeated as part of an optimisation procedure in the search for the best alignment. The criterion for optimality is the minimization of the weighted sum of the volumes of each descriptor type.

The fitting procedure utilizes the current conformations of the molecules and works as follows. For each atom in the molecule undergoing refitting, the nearest atom of the anchored ligand is found. This elicits an atom–atom correspondence list from which a least-square superposition matrix is determined and the molecule to be refit is transformed to a new position and orientation. After this step, the nearest anchored ligand atom to a particular atom of the refit molecule may be different from the one previously identified. The correspondences are, thus, re-assigned and the refitting is performed again. This process is repeated until convergence is reached or a maximum of ten iterations have been conducted.

The scoring function that assesses the quality of the alignment comprises two terms. First, the weighted sum of volumes of descriptor types across all the current structures is calculated. This is conducted using a grid-based protocol. Each descriptor point is surrounded by a sphere of an appropriate radius assigned from the bond radius of the corresponding atom. The ligands are surrounded by a gridded box and, for each descriptor type, counts are obtained of the number of grid points within the spherical radii of the corresponding feature points. The weighted sum of descriptor type volumes yields the volume of the molecular alignment. The second term comprises intramolecular steric clash penalties. If the distance between two atoms is less than 70% of the sum of van

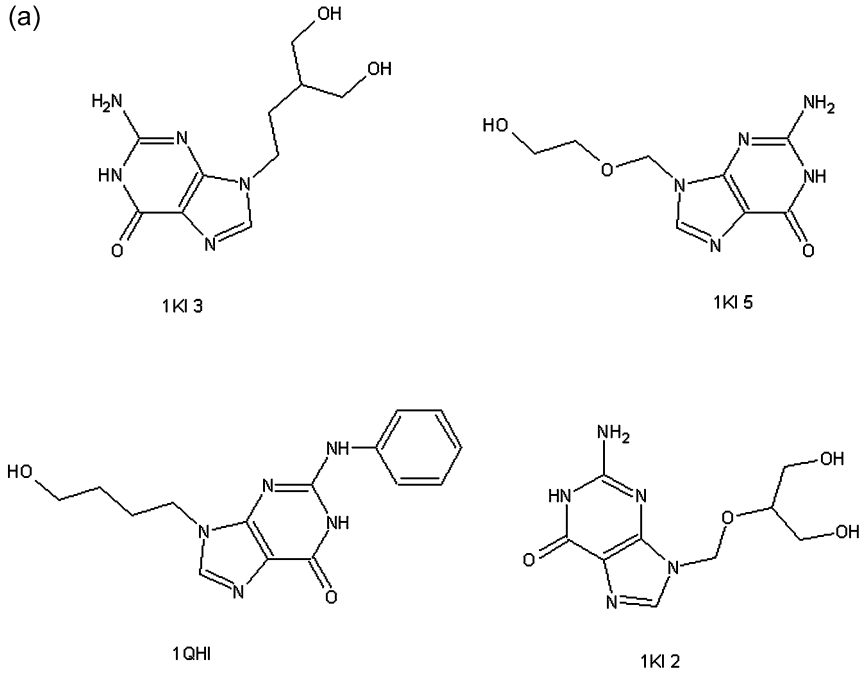


*Figure 1.* Chemical structures of all ligands considered for the derivation of binding site models: (a) HSV-1 and (b) PARP.

der Waals radii, the amount by which the distance is lower than that value is added to the score as a penalty.

A simulated annealing [37] optimisation procedure is used to find the best superposition with the multiple structures occupying a minimum volume. This procedure initially sets the annealing temperature and invokes system transitions in an iterative manner, which are either accepted or rejected according to the Metropolis condition [38]. The system is gradually cooled during the simulation as the minimum structure is approached. Several types of transition are used to generate new configurations: (i) rigid body translations, (ii) rotations, and (iii) conformational changes of a randomly selected ligand. Transition type (iii) allows conformational flexibility and multiple ligand states to be handled efficiently. Each ligand can be represented by several conformers and/or tautomers and protonation/ionization states, which are generated prior to the superposition process. When transition type (iii) is selected, one of these states is randomly chosen and its conformation is modified by applying a random rotation around a selected rotatable bond. If the transition is accepted the conformation is stored and used in the next transition involving the same ligand. Consequently, ligand flexibility is incorporated by using static conformers/states (which are generated prior to the superposition process as input) and dynamic conformational changes during the simulation. This method, along with its predecessor SLATE for pairwise superposition [39, 40], allows regions of partial similarity to be efficiently matched within a set of molecular structures [35].Due to the stochastic nature of the simulated annealing process, each optimisation can produce a different alignment. Multiple alignments so generated are ranked and clustered before selecting one or more models. The full set of interaction features associated with an aligned set of molecules comprises a binding site model. These features may be clustered in three-dimensional space to obtain a reduced set representation, or binding site model [35].

It is important to realize that there are a number of different ways in which to perform and achieve suitable alignments from sets of molecules [41–47], and it would be relatively straightforward to derive a suitable binding site model once a successful molecular alignment has been obtained by any of these or, indeed, other appropriate superposition/alignment methods.

We duly applied the above superposition method to sets of ligands known to bind to the two proteins detailed earlier. In the case of HSV-1, we extracted ligands from four of the PDB files listed in Table 1 (1ki2, 1ki3, 1ki5 and 1qhi). Their 2D chemical structures can be seen in Figure 1a. The ligand from 1qhi

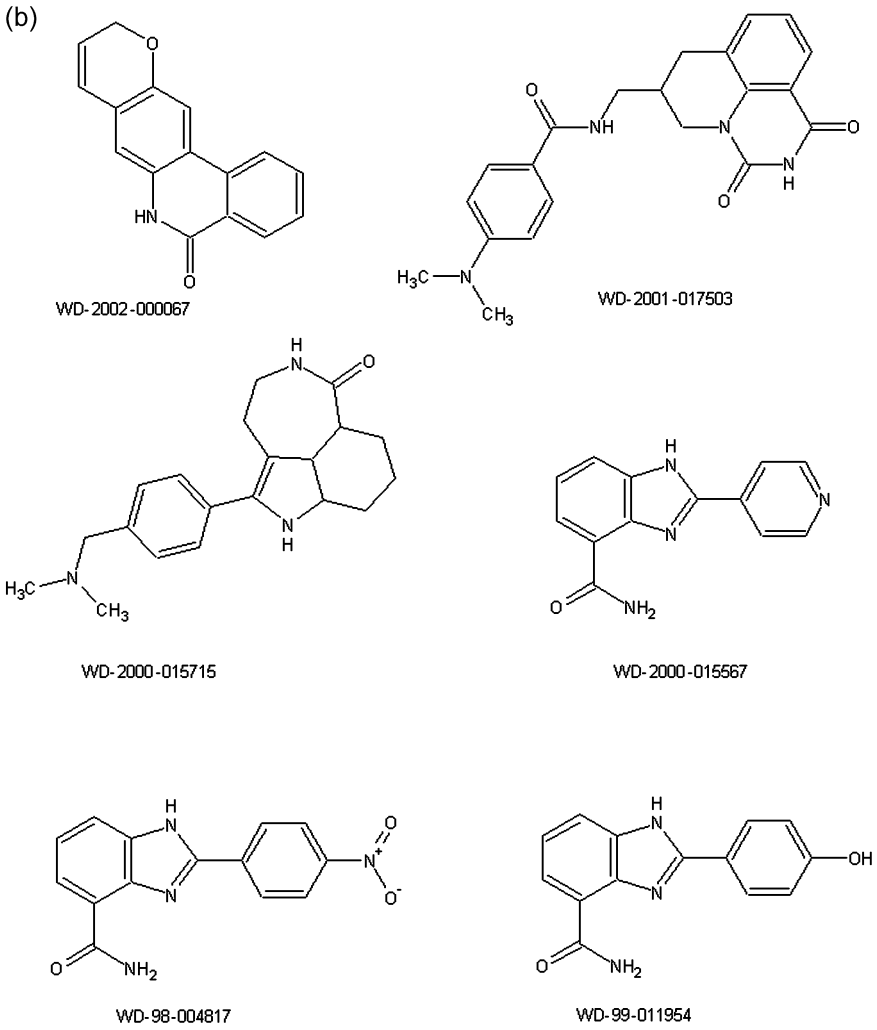


*Figure 1.* (Continued).

was used as a template, and the positions and conformations of all the other ligands were randomized. Using our ligand superposition method, we then optimized the molecular alignments on the basis of their interaction features. The resulting ligand superposition and derived binding site model were then overlayed into an appropriate crystal structure. For HSV-1, 1qhi was chosen so that the frame of reference would be that of the frozen co-crystallized ligand. The binding site model for HSV-1 can be seen in Figure 2b. This kind of binding site model is comprised of the full set of features associated with the aligned set of ligands.

In the case of PARP, a different approach was followed. A set of six known non-co-crystallized ligands[48] (WD-2002-000067, WD-2001-017503, WD-2000-015715, WD-2000-015567, WD-98-004817 and WD-99-011954) was assembled and Quasi2 was employed to optimise the superposition of their interaction features while allowing all ligands full conformational flexibility. The 2D chemical structures of all the ligands can be seen in Figure 1b. A binding site model was derived from these ligands in the same way as described above for HSV-1, but without a template reference ligand. The resulting ligand superposition and derived binding site model (which can be seen in Figure 2d) were overlayed into the 2pax crystal structure, since the ligand present in this structure has a common benzamide molecular core with one of the ligands used in the superposition.

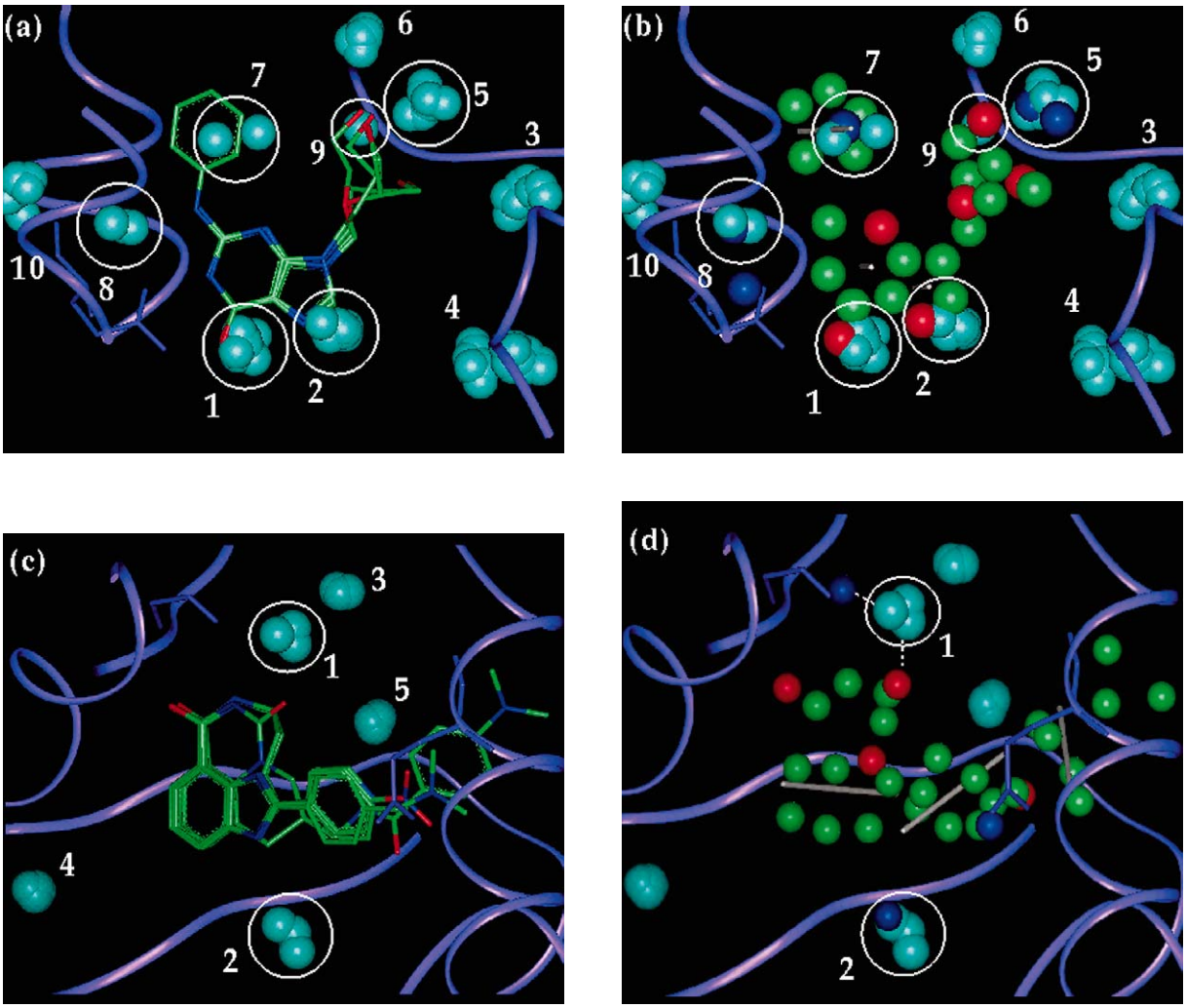


*Figure 2.* Prediction of tightly bound water molecules (shown as spheres in cyan) and binding site model (shown as spheres in green for steric features, red for ligand hydrogen-bond acceptor group and blue for ligand hydrogen-bond donor group; aromatic vectors are shown as grey sticks). The white circles highlight water clusters (labeled according to Tables 1 and 2) that match some of the binding site model features. (a) Tightly bound water clusters found in HSV-1, with aligned ligands shown for reference. (b) Binding site model for HSV-1. (c) Tightly bound water clusters found in PARP, with aligned ligands shown for reference. (d) Binding site model for PARP.

*Matching site model points to binding site features*

A site model is comprised of up to four feature types – steric, aromatic, ligand acceptors and donor projections – graphically represented as green, stick, red and blue features respectively, cf. Figure 2. The matching of site model features to binding site features is facilitated by reading the site model into the frame of reference of the protein structure – coincident features are readily visually identified through examination of the resultant overlay. Ideally, site features and model features will be coincident within a reasonable distance – such comparisons are generally employed in the validation of ligand superposition methods. In cases where features are not coincident, the quality of the site model must be questioned or non-coincident features satisfactorily explained from our knowledge of the system being studied.

*Prediction of tightly bound water molecules*

A multivariate logistic method for identifying bound water molecules on the surface of a binding site, named WaterScore, has been recently introduced and validated [31]. This method is based on a survey of a diverse set of proteins in their free and liganded

states, where water molecules in the binding site of these proteins were classified as displaceable (if no matching water molecules were found between the free and liganded forms) and bound (if a match was found). The subset of bound water molecules was thus defined as that containing tightly bound water molecules that remained in the same position upon ligand binding. A number of structural properties (B-factors, solvent-contact surface area, number of protein atomic contacts and the total hydrogen bond energy) for each water molecule in the above two subsets were computed and the best multivariate logistic regression models were chosen to discriminate between bound and displaceable water molecules in the form of a response variable that corresponds to the probability of the higher value (which, by definition, takes a value of 1 and was chosen to represent water molecules in their bound 'state').

The best logistic model that was obtained contained only three of the above variables and had a 70% level of significance:

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

with

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

where Bf stands for B-factor, SCSA for solvent-contact surface area and NPAC for number of protein atomic contacts. 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 and $b_3$ = 84.458. The lower limit for P(Y=1) for bound water molecules was found to be higher than $10^{-20}$ for an appropriate discrimination between bound and displaceable water molecules. Full details of this method, the multivariate logistic regression approach used and the validation of the best models can be found elsewhere [31].

We applied this method to the various crystal structures of HSV-1 and PARP. For each of these crystal structures listed in Tables 1 and 2, only water molecules within a cutoff distance of 7.0 Å from any ligand atom were considered. Isotropic temperature B-factors were read directly from the PDB files of the proteins considered. The solvent-contact surface area for each water molecule was computed using the program NACCESS 2.1.1 [49], which calculates the atomic contact surface defined by rolling a probe of a given size around the van der Waals surface and following the coordinates of the surface of the probe [50]. The radius of the rolling probe used was 1.2 Å,

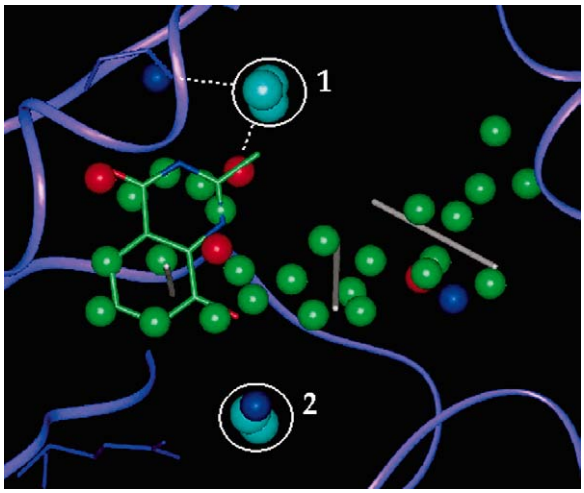


*Figure 3.* Co-crystallized ligand (in sticks) found in 4pax (PARP), binding site model (shown as spheres in green for steric features, red for ligand hydrogen-bond acceptor group and blue for ligand hydrogen-bond donor group; aromatic vectors are shown as grey sticks) and water clusters 1 and 2 (highlighted with white circles). The two relevant interaction features for cluster 1 are highlighted by dashed lines. The hydrogen-bond donor interaction feature (in blue) in cluster 2 is already enclosed by the white circle.

which has been used elsewhere for the exploration of protein surfaces [51]. The number of atomic contacts with protein atoms was calculated using a cut-off of 3.5 Å from the centre of each water molecule. A full listing of the values of these properties and the calculated probabilities (P(Y)) for each water molecule in the clusters listed in Tables 1 and 2 can be found in Tables 3 and 4.

All crystal structures for each kind of protein were then overlayed with the program Swiss PdbViewer [52] by considering only those protein atoms within 7.0 Å of any ligand atom. All water molecules, in their new frame of reference, were then inspected visually to identify relevant clusters of water molecules occupying equivalent positions in different crystal structures. Water molecules in those crystal structures that seemed to have been sterically displaced by one or more atoms a ligand in another crystal structure were classified as sterically displaced by a ligand. In some cases water molecules were not found in certain crystal structures because of a sidechain reorientation that displaced them, in these cases; the relevant sidechain was identified. In all the crystal structures analyzed, only two water molecules belonging to a relevant cluster were not predicted to be bound water molecules, and were thus classified as displaceable (see Table 1). Each column in Tables 1 and 2 lists all

*Table 1.* List of tightly bound water molecules in HSV-1 crystal structures.

| PDB code | R[a] (Å) | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 | Cluster 9 | Cluster 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1ki2 | 2.2 | SD[b] | SD | Not found | 532 | SD | Not found | 695 | SD | SD | Gln 125 |
| 1ki3 | 2.4 | SD | SD | Gln 221 | 507 | SD | Glu 83 | 542 | SD | SD | 559 |
| 1ki4 | 2.3 | 572 | 568 | 712 | 576 | 570 | Glu 83 | SD | Gln 125 | SD | 525 |
| 1ki6 | 2.4 | 566 | 564 | 644 | 570 | 775 | 562 | SD | Gln 125 | SD | 524 |
| 1ki7 | 2.2 | 551 | 521 | Arg 220 | 559 | Arg 163 | Glu 83 | SD | Gln 125 | SD | 582 |
| 1ki8 | 2.2 | 532 | 515 | 710 | 554 | 569 | Glu 83 | SD | Gln 125 | SD | 574 |
| 1kim | 2.1 | 435 | 433 | 459 | 437 | 434 | Glu 83 | SD | Gln 125 | SD | 413 |
| 1qhi | 1.9 | SD | SD | 523 | 599 | 763 | 503 | SD | 521 | SD | 550 |
| 2ki5 | 1.9 | SD | SD | 592 | 598 | 776 | 587 | D | 586 | 779 | 545 |
| 1e2h | 1.9 | 47 | 28 | 92 | 100 | SD | 14 | Not found | 46 | Not found | 45 |
| 1e2i | 1.9 | 38 | 27 | 81 | 28 | SD | 14 | SD | 37 | SD | 36 |
| 1e2j | 2.5 | 33 | 17 | 46 | Not found | SD | 10 | SD | SD | SD | 24 |
| 1e2k | 1.7 | 64 | 40 | 122 | 7 | Not found | 17 | SD | Gln 125 | SD | 63 |
| 1e2l | 2.4 | 29 | Not found | Arg 222 | 60 | SD | 14 | SD | Gln 125 | SD | 28 |
| 1e2m | 2.2 | 51 | 39 | Not found | Not found | Arg 226 | 21 | SD | Gln 125 | SD | 50 |
| 1e2n | 2.2 | 39 | 17 | 56 | D[c] | SD | 35 | SD | Gln 125 | SD | Gln 125 |

[a]R = resolution factor. [b]SD = sterically displaced by the ligand. [c]D = displaceable.

*Table 2.* List of tightly bound water molecules in PARP crystal structures.

| PDB code | R[a] (Å) | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|---|
| 1efy | 2.2 | 107 | 52 | 132 | 108 | 8 |
| 1pax | 2.4 | 82 | 65 | 5 | 9 | 8 |
| 2pax | 2.4 | SD | SD[b] | 19 | 9 | 4 |
| 3pax | 2.4 | 75 | 61 | 4 | 9 | 7 |
| 4pax | 2.8 | SD | Not found | 5 | 9 | 8 |

[a]R = resolution factor. [b]SD = sterically displaced by the ligand.

bound water molecules (by their original numbers in their corresponding PDB files) pertaining to the same cluster.

## Results and discussion

### *Comparison of binding site models with predicted bound water molecules*

In the case of HSV-1, a good overlay of all the tightly bound water molecules was obtained with the various crystal structures analyzed. Figure 2a shows the corresponding superposition of all the predicted bound water molecules (with the ligands described below shown as reference) and their arrangement into the clusters outlined in Table 1. Each crystal structure tends to have a tightly bound water molecule in the same position for all the nine clusters found, adding credibility to the lists of bound water molecules predicted for each crystal structure independently. There were instances where the ligand occupied one of these positions and, obviously, the corresponding water molecule would have been sterically displaced from the binding site (clusters 1, 2, 5, 7, 8 and 9). In some clusters, no water molecules were found in some of the crystal structures, most likely due to the fact that it was not possible to resolve those water-molecule positions when the crystal structure was determined. This illustrates the fact that, whenever possible, working with a series of ligand–protein complexes can be as important as making an accurate prediction of tightly bound water molecules, since the joint analysis of several crystal structures can help to identify missing water molecules.

HSV-1 exhibits some degree of rotameric flexibility in its binding site and Table 1 identifies also those sidechains that have putatively displaced water molecules as a result of these conformational changes. In only one case (not shown), observed in the 1e2j crystal structure, a tightly bound water molecule (number

*Table 3.* Structural properties and predicted probabilities for the water molecules found in the selected clusters of HSV-1 crystal structures.

| PDB code | Cluster | Water ID | *Bf* | *SCSA* | NPAC | P(Y) |
|---|---|---|---|---|---|---|
| 1ki2 | 4 | 532 A | 19.05 | 0.000 | 1 | 1 |
| | 7 | 695 A | 18.54 | 2.863 | 1 | 1 |
| 1ki3 | 4 | 507 A | 19.84 | 0.000 | 2 | 1 |
| | 7 | 542 A | 16.62 | 4.311 | 1 | 1 |
| | 10 | 559 A | 28.02 | 1.528 | 3 | 1 |
| 1ki4 | 1 | 572 A | 18.26 | 0.854 | 4 | 1 |
| | 2 | 568 A | 14.64 | 1.516 | 5 | 1 |
| | 3 | 712 A | 45.02 | 0.302 | 3 | 1 |
| | 4 | 576 A | 12.55 | 0.000 | 3 | 1 |
| | 5 | 570 A | 15.49 | 2.473 | 4 | 1 |
| | 10 | 525 A | 24.34 | 0.016 | 5 | 1 |
| 1ki6 | 1 | 566 A | 4.67 | 0,721 | 2 | 1 |
| | 2 | 564 A | 15.13 | 1.116 | 5 | 1 |
| | 3 | 644 A | 21.52 | 0.234 | 4 | 1 |
| | 4 | 570 A | 8.93 | 0.000 | 3 | 1 |
| | 5 | 775 A | 27.52 | 0.003 | 8 | 1 |
| | 6 | 562 A | 19.56 | 0.358 | 4 | 1 |
| | 10 | 524 A | 25.01 | 0.053 | 4 | 1 |
| 1ki7 | 1 | 551 A | 6.23 | 0.593 | 3 | 1 |
| | 2 | 521 A | 2.93 | 1.515 | 3 | 1 |
| | 4 | 559 A | 7.23 | 0.000 | 2 | 1 |
| | 10 | 582 A | 27.48 | 0.064 | 3 | 1 |
| 1ki8 | 1 | 532 A | 17.45 | 0.693 | 3 | 1 |
| | 2 | 515 A | 9.26 | 1.821 | 3 | 1 |
| | 3 | 710 A | 61.27 | 0.156 | 8 | 1 |
| | 4 | 554 A | 17.59 | 0.214 | 1 | 1 |
| | 5 | 569 A | 22.01 | 0.965 | 5 | 1 |
| | 10 | 574 A | 18.08 | 0.132 | 3 | 1 |
| 1kim | 1 | 435 A | 11.56 | 0.794 | 3 | 1 |
| | 2 | 433 A | 14.72 | 1.777 | 3 | 1 |
| | 3 | 459 A | 22.69 | 0.000 | 6 | 1 |
| | 4 | 437 A | 22.87 | 0.000 | 2 | 1 |
| | 5 | 434 A | 18.02 | 2.168 | 5 | 1 |
| | 10 | 413 A | 14.05 | 0.038 | 4 | 1 |
| 1qhi | 3 | 523 | 28.18 | 0.009 | 4 | 1 |
| | 4 | 599 | 22.63 | 0.228 | 2 | 1 |
| | 5 | 763 | 18.51 | 1.082 | 5 | 1 |
| | 6 | 503 | 15.38 | 0.195 | 6 | 1 |
| | 8 | 521 | 9.83 | 1.427 | 2 | 1 |
| | 10 | 550 | 12.97 | 0.002 | 3 | 1 |
| 2ki5 | 3 | 592 | 27.94 | 0.449 | 4 | 1 |
| | 4 | 598 | 26.55 | 0.063 | 3 | 1 |
| | 5 | 776 | 39.85 | 0.358 | 4 | 1 |
| | 6 | 587 | 18.38 | 0.000 | 5 | 1 |
| | 8 | 586 | 22.05 | 0.910 | 6 | 1 |
| | 9 | 779 A | 21.33 | 4.932 | 1 | 1 |
| | 10 | 545 | 21.85 | 0.047 | 2 | 1 |
| 1e2h | 1 | 47 Z | 27.57 | 0.189 | 5 | 1 |
| | 2 | 28 Z | 34.30 | 1.456 | 3 | 1 |
| | 3 | 92 Z | 45.20 | 0.027 | 4 | 1 |
| | 4 | 100 Z | 37.75 | 0.027 | 3 | 1 |
| | 6 | 14 Z | 39.19 | 0.569 | 5 | 1 |
| | 8 | 46 Z | 30.69 | 0.000 | 7 | 1 |
| | 10 | 45 Z | 28.40 | 0.000 | 3 | 1 |
| 1e2i | 1 | 38 Z | 38.73 | 0.460 | 5 | 1 |
| | 2 | 27 Z | 44.25 | 2.810 | 3 | 1 |
| | 3 | 81 Z | 33.51 | 0.000 | 4 | 1 |
| | 4 | 28 Z | 45.58 | 0.000 | 3 | 1 |
| | 6 | 14 Z | 38.01 | 0.360 | 4 | 1 |
| | 8 | 37 Z | 37.40 | 0.344 | 8 | 1 |
| | 10 | 36 Z | 41.66 | 0.819 | 4 | 1 |
| 1e2j | 1 | 33 Z | 15.49 | 0.930 | 2 | 1 |
| | 2 | 17 Z | 42.45 | 1.524 | 4 | 1 |
| | 3 | 46 Z | 51.29 | 0.124 | 3 | 1 |
| | 6 | 10 Z | 13.71 | 1.366 | 4 | 1 |
| | 10 | 24 Z | 29.35 | 1.266 | 3 | 1 |
| 1e2k | 1 | 64 Z | 21.75 | 0.809 | 3 | 1 |
| | 2 | 40 Z | 21.45 | 2.418 | 4 | 1 |
| | 3 | 122 Z | 38.26 | 0.846 | 4 | 1 |
| | 4 | 7 Z | 27.44 | 0.000 | 2 | 1 |
| | 6 | 17 Z | 28.06 | 0.467 | 4 | 1 |
| | 10 | 63 Z | 29.40 | 1.789 | 3 | 1 |
| 1e2l | 1 | 29 Z | 33.96 | 1.532 | 3 | 1 |
| | 4 | 60 Z | 37.13 | 0.035 | 1 | $5.4 \times 10^{-16}$ |
| | 6 | 14 Z | 34.17 | 0.810 | 3 | 1 |
| | 10 | 28 Z | 32.15 | 0.214 | 2 | 1 |
| 1e2m | 1 | 51 Z | 18.39 | 0.486 | 3 | 1 |
| | 2 | 39 Z | 32.18 | 0.748 | 4 | 1 |
| | 6 | 21 Z | 27.78 | 1.693 | 5 | 1 |
| | 10 | 50 Z | 20.26 | 0.003 | 5 | 1 |
| 1e2n | 1 | 39 Z | 37.09 | 0.521 | 4 | 1 |
| | 2 | 17 Z | 40.17 | 2.445 | 3 | 1 |
| | 3 | 56 Z | 60.55 | 1.561 | 3 | 1 |
| | 4 | 9 Z | 76.16 | 0.109 | 1 | $1.6 \times 10^{-105}$ |
| | 6 | 35 Z | 50.64 | 3.819 | 6 | 1 |

25) was seen to occupy the position left vacant after the rotameric movement of Gln 125, and matching the hydrogen-bond acceptor group shown in Figure 2b below cluster 6 (see below). Water molecules 38 and 85 from 1e2m and water molecule 7 from 1e2l were also predicted to be tightly bound, but are not shown for reasons of clarity in Figure 2a, as they do not belong

*Table 4.* Structural properties and predicted probabilities for the water molecules found in the selected clusters of PARP crystal structures.

| PDB code | Cluster | Water ID | *Bf* | *SCSA* | NPAC | P(Y) |
|---|---|---|---|---|---|---|
| 1efy | 1 | 52 | 28.66 | 6.542 | 4 | 1 |
| | 2 | 107 | 32.19 | 9.199 | 1 | $2.7 \times 10^{-13}$ |
| | 3 | 132 | 21.39 | 3.732 | 2 | 1 |
| | 4 | 108 | 28.00 | 1.761 | 1 | 1 |
| | 5 | 8 | 17.27 | 0.000 | 6 | 1 |
| 1pax | 1 | 65 | 51.69 | 5.578 | 4 | 1 |
| | 2 | 82 | 71.58 | 8.540 | 4 | 1 |
| | 3 | 5 | 29.45 | 4.145 | 2 | 1 |
| | 4 | 9 | 31.43 | 4.861 | 1 | $1.8 \times 10^{-7}$ |
| | 5 | 8 | 20.11 | 0.000 | 5 | 1 |
| 2pax | 3 | 19 | 29.45 | 0.912 | 2 | 1 |
| | 4 | 9 | 32.02 | 4.422 | 2 | 1 |
| | 5 | 4 | 22.35 | 0.000 | 6 | 1 |
| 3pax | 1 | 61 | 61.99 | 6.340 | 5 | 1 |
| | 2 | 75 | 59.67 | 7.975 | 3 | $8.3 \times 10^{-2}$ |
| | 3 | 4 | 27.17 | 5.300 | 2 | 1 |
| | 4 | 9 | 29.38 | 5.105 | 1 | $5.1 \times 10^{-1}$ |
| | 5 | 7 | 19.68 | 0.000 | 6 | 1 |
| 4pax | 3 | 5 | 25.82 | 6.394 | 2 | 1 |
| | 4 | 9 | 27.80 | 5.429 | 1 | $9.2 \times 10^{-2}$ |
| | 5 | 8 | 16.48 | 0.000 | 2 | 1 |

to any identified cluster and do not occupy relevant positions in the binding site.

In the case of PARP, a good overlay of all the predicted tightly bound water molecules was also obtained. Figure 2c shows the resulting superposition of significant bound water molecules (with the ligands described below shown as reference) and how they fall into clusters 1 and 2 outlined in Table 2 (the other clusters are too far from the ligands and are not shown for clarity). It can be seen that each crystal structure has a tightly bound water molecule in the same position for all the five clusters found, once again adding credibility to the independent predictions made on each crystal structure. In the case of 2pax, no water molecules were found for clusters 1 and 2 because the ligand molecule occupies those positions and hence the water molecules are likely to have been sterically displaced. In the case of 4pax, the same can be said about the absence of a water molecule in cluster 1. For this protein, we could not find a water molecule that would match cluster 2, but, as before, we ascribe the absence of this water molecule to the fact that this is the crystal structure with the lowest resolution (2.8 Å) and, consequently, it may not have been possible to resolve this water-molecule position. As before, this example illustrates the benefit of working with a series of ligand–protein complexes when identifying the positions of tightly bound water molecules.

*Integrating results of water analyses to binding site model interpretation*

*Thymidine kinase.* As detailed above, some of the ligand-derived binding site model features can be immediately visually matched to corresponding hydrogen bonding groups in the binding site of the proteins. However, a number of features do not have matching sitepoint groups in the protein. A closer inspection reveals that their matching partner groups are indeed the water molecules that we have earlier predicted as being tightly bound to the surface of the binding sites. These non-matched protein sitepoint positions are highlighted by white circles in the binding site model and clearly match the water clusters highlighted in Figure 2a. Clusters 1, 2, 5, 7, 8 and 9 contain water molecules that may or may not be sterically displaced by a ligand. We also conclude that the water molecules in cluster 6 do not always act as hydrogen-bond donors, as depicted in the binding site model in Figure 2b: in the 1e2j crystal structure, the water molecule in that position is observed to act as a hydrogen-bond acceptor.

*Poly ADP ribose polymerase.* As for thymidine kinase, an examination of the PARP site model highlights that a number of the features associated with the aligned set of ligands can be matched to corresponding hydrogen bonding groups in the protein. We again find that those features that did not have matching protein receptor sitepoint groups in the protein do indeed match water molecules that we had earlier predicted as being tightly bound to the binding site. As before, these non-matched protein sitepoint group positions are highlighted by white circles in the binding site model and again match the water clusters highlighted in Figure 2c. Clusters 1 and 2 contain water molecules that may or may not be sterically displaced by a ligand. Clusters 3 and 5 are further away from the ligands than what may seem from inspecting Figure 2c owing to the perspective chosen to best depict the binding site in these images.

Specifically illustrated by our concurrent analysis of site model features and bound water predictions for PARP, when potent non-co-crystallized ligands are

used to derive a binding site model and the resultant model is used in conjunction with structural information as illustrated in Figure 3, we can immediately see how the combination of ligand and structure-based design methodologies is beneficial. In Figure 3 we have illustrated the co-crystallized ligand from 4pax (stick rendered). From our analyses of bound water molecules, for this particular structure neither cluster 1 nor cluster 2 are present. As a consequence, were this crystal structure used in isolation – as is frequently the case in the early stages of a design program – no consideration would be given to the role of these water molecules in subsequent drug design strategies. However, our water analysis indicates the likelihood of bound water molecules in these regions and, further to this, our ligand-derived pharmacophoric binding site model indicates that known potent PARP ligands may interact with these predicted bound water molecules through the interaction features highlighted in Figure 3. Additionally, the ligand co-crystallized in 4pax is relatively small in size – the nature of the binding site in PARP is such that many potential directional vectors could be explored in designing novel modulators. Taking the result of our pharmacophore analysis into the binding site crystal structure frame of reference immediately indicates an appropriate extension vector direction for exploration. Combining the knowledge gained from applying bound water molecule conservation analysis and ligand-based pharmacophores together with available crystal structure data for a target binding site can underpin and advance consensus drug design approaches, as each method brings exploitable and relevant information to the drug design process.

## Conclusions

The two ligand–protein systems that we have analyzed exemplify the importance of combining a structure-based approach to the analysis of the binding site of a protein with the ligand-based approach of deriving a binding site model from the optimum alignment of the interaction features of known ligands. In particular, the agreement of these two methods on the prediction and role of tightly bound water molecules illustrates how important water molecules can be in participating directly in the interactions between ligands and their protein receptors.

More important is our intention to highlight the role that water molecules can have when interpreting or using a pharmacophore derived from a set of known active ligands. This work demonstrates that pharmacophore-derived binding site models extrapolated from available ligand information can often include features arising from interactions with tightly bound water molecules and not solely protein receptor sitepoints. Both structure- and ligand-based drug design approaches should be applied in conjunction whenever possible (consensus drug design) and the inclusion of tightly bound water molecules should be always considered when rationalizing the properties of a binding site.

## Acknowledgements

D.G.L. gratefully acknowledges the contribution of Dr. Paulette Greenidge to the development of the concept of consensus design. A.T.G.S. would like to thank Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico for the award of a postgraduate scholarship and Universities UK for an Overseas Research Scheme award. R.L.M. is also a Research Fellow of Hughes Hall, Cambridge, UK.

## References

1. Takano, K., Yamagata, Y. and Yutani, K., Protein Eng., 16 (2003) 5.
2. Davis, A.M., Teague, S.J. and Kleywegt, G.J., Angew. Chem. Int. Ed. Engl., 42 (2003) 2718.
3. Engh, R.A., Brandstetter, H., Sucher, G., Eichinger, A., Baumann, U., Bode, W., Huber, R., Poll, T., Rudolph, R. and von der Saal, W., Structure, 4 (1996) 1353.
4. Rejto, P.A. and Verkhivker, G.M., Proteins Struct. Funct. Genet., 28 (1997) 313.
5. Finley, J.B., Atigadda, V.R., Duarte, F., Zhao, J.J., Brouillette, W.J., Air, G.M. and Luo, M., J. Mol. Biol., 293 (1999) 1107.
6. Palomer, A., Pérez, J.J., Navea, S., Llorens, O., Pascual, J., García, Ll. and Mauleón, D., J. Med. Chem., 43 (2000) 2280.
7. Poornima, C.S. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 500.
8. Marrone, T.J., Briggs, J.M. and McCammon, J.A., Ann. Rev. Pharmacol. Toxicol., 37 (1997) 71.
9. 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. and Ericksonviitanen, S., Science, 263 (1994) 380.
10. Chen, J.M., Xu, S.L., Wawrzak, Z., Basarab, G.S. and Jordan, D.B., Biochemistry, 37 (1998) 17735.
11. Mikol, V., Papageorgiou, C. and Borer, X., J. Med. Chem., 38 (1995) 3361.
12. Holdgate, G.A., Tunnicliffe, A., Ward, W.H.J., Weston, S.A., Rosenbrock, G., Barth, P.T., Taylor, I.W.F., Pauptit, R.A. and Timms, D., Biochemistry, 36 (1997) 9663.
13. Cherbavaz, D.B., Lee, M.E., Stroud, R.M. and Koschl, D.E., J. Mol. Biol., 295 (2000) 377.

14. Rarey, M., Kramer, B. and Lengauer, T., Proteins Struct. Funct. Genet., 34 (1998) 17.
15. Schnecke, V. and Kuhn, L.A., Perspect. Drug Discov. Des., 20 (2000) 171.
16. Pospisil, P., Kuoni, T., Scapozza, L. and Folkers, G., J. Recept. Signal Transduct. Res., 22 (2002) 141.
17. Pastor, M., Cruciani, G. and Watson, K.A., J. Med. Chem., 40 (1997) 4089.
18. Mancera, R.L., J. Comput.-Aided Mol. Design, 16 (2002) 479.
19. Nakasako, M., J. Mol. Biol., 289 (1999) 547.
20. Faerman, C.H. and Karplus, P.A., Proteins Struct. Funct. Genet., 23 (1995) 1.
21. Schwabe, J.W.R., Curr. Opin. Struct. Biol., 7 (1997) 126.
22. Carrell, H.L., Glusker, J.P., Burger, V., Manfre, F., Tritsch, D. and Biellmann, J.-F., Proc. Natl. Acad. Sci. USA, 86 (1989) 4440.
23. Chung, E., Henriques, D., Renzoni, D., Zvelebil, M., Bradshaw, J.M., Waksman, G., Robinson, C.V. and Ladbury, J.E., Struct. Fold. Des., 6 (1998) 1141.
24. Ogata, K. and Wodak, S.J., Protein Eng., 15 (2002) 697.
25. Loris, R., Langhorst, U., De Vos, S., Decanniere, K., Bouckaert, J., Maes, D., Transue, T.R. and Steyaert, J., Proteins Struct. Funct. Genet., 36 (1999) 117.
26. Loris, R., Stas, P.P.G. and Wyns, L., J. Biol. Chem., 269 (1994) 26722.
27. Poornima, C.S. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 521.
28. Ehrlich, L., Reckzo, M., Bohr, H. and Wade, R.C., Protein Eng., 11 (1998) 11.
29. Raymer, M.L., Sanschagrin, P.C., Punch, W.F., Venkataraman, S., Goodman, E.D. and Kuhn, L.A., J. Mol. Biol., 265 (1997) 445.
30. Sanschagrin, P.C. and Kuhn, L.A., Protein Sci., 7 (1998) 2054.
31. García-Sosa, A.T., Mancera, R.L. and Dean, P.M., J. Mol. Model., 9 (2003) 172.
32. Anstead, G.M., Carlson, K.E. and Katzenellenbogen, J.A., Steroids, 62 (1997) 268.
33. Grünenberg, S., Stubbs, M.T. and Klebe, G., J. Med. Chem., 45 (2002) 3588.
34. Brenk, R., Naerum, L., Grädler, U., Gerber, H-D., Garcia, G.A., Reuter, K., Stubbs, M.T. and Klebe, G., J. Med. Chem., 46 (2003) 1133.
35. Perry, N.C., Lloyd, D.G., Todorov, N.P. and Alberts, I.L. In: QSAR Proceedings 2002: EuroQSAR 2002: Designing Drugs and Crop Protectants. Blackwell Publishing, Oxford, UK, 2003, pp. 68–72.
36. Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 10 (1996) 607.
37. Kirkpatrick, S., Gellat C.D. and Vecchi, M.P., Science, 220 (1983) 671.
38. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E., J. Chem. Phys., 21 (1953) 1087.
39. Mills, J.E.J., De Esch, I.J., Perkins, T.D.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 1 (2001) 81.
40. De Esch, I.J., Mills, J.E.J., Perkins, T.D.J., Romeo, G., Hoffmann, M., Wieland, K., Leurs, R., Menge, W.M.P.B., Nederkoorn, P.H.J., Dean, P.M. and Timmerman, H., J. Med. Chem., 44 (2001) 1666.
41. Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7 (1993) 83.
42. McMartin, C. and Bohacek, R.S., J. Comput.-Aided Mol. Design, 9 (1995) 237.
43. Handschuh, S., Wagener, M. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 38 (1998) 220.
44. Lemmen, C., Lengauer, T. and Klebe, G., J. Med. Chem., 41 (1998) 4502.
45. Lemmen, C., Hiller, C. and Lengauer, T., J. Comput.-Aided Mol. Design, 12 (1998) 491.
46. Goldman, B.B. and Wipke, W.T., J. Chem. Inf. Comput. Sci., 40 (2000) 644.
47. Melani, F., Graterri, P., Adamo, M. and Bonaccini, C., J. Med. Chem., 46 (2003) 1359.
48. Derwent World Drug Alerts, Derwent Information, London, 2002.
49. Hubbard, S.J. and Argos, P., Protein Eng., 8 (1995) 1011.
50. Lee, B. and Richards, F.M., J. Mol. Biol., 55 (1971) 379.
51. Poornima, C.S. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 513.
52. Guex, N. and Peitsch, M.C., Electrophoresis, 18 (1997) 2714.