Ligand−Protein Cross-Docking with Water Molecules

Ramasamy Thilagavathi[§] and Ricardo L. Mancera*

Curtin Health Innovation Research Institute, Western Australian Biomedical Research Institute, School of Pharmacy and Biomedical Sciences, Curtin University of Technology, GPO Box U1987, Perth WA 6845, Australia

Received September 12, 2009

The accuracy of ligand−protein docking may be affected by the presence of water molecules on the surface of the protein. Cross-docking simulations have been performed on a number of ligand−protein complexes for various proteins whose crystal structures contain water molecules in their binding sites. Only common sets of water molecules found in the binding site of the proteins were considered. A statistically significant overall increase in accuracy was observed when water molecules were included in cross-docking simulations. These results confirm the importance of including water molecules whenever possible in ligand−protein docking simulations.

INTRODUCTION

Cross-docking involves docking several ligands found in various crystal structures of the same protein to a single rigid protein conformation. This can result in significant errors if an induced fit conformational change in the protein upon ligand binding is necessary. Several approaches have been reported to address this protein flexibility issue during molecular docking. For example, GOLD[1] performs automated docking with full ligand flexibility and partial protein flexibility. FlexE[2] uses multiple experimentally obtained protein structures. IFREDA[3] accounts for protein flexibility by generating a discrete set of receptor conformations. Glide[4] performs a complete systematic search of the conformational, orientational, and positional space of the docked ligand. In addition, protein flexibility in ligand−protein docking has also been modeled by explicit solvent molecular dynamics simulations.[5−8]

Incorporating hydration effects during molecular docking is another challenging problem. Three main computational approaches can be used to model hydration, with different levels of accuracy.[9] Free energy methods, such as free energy perturbation (FEP) and thermodynamic integration, are the most accurate techniques used for predicting changes to the free energies of ligand binding upon removal/addition of specific water molecules. Computational simulations have been used successfully to compute the free energy changes associated with hydrating binding sites and with displacing water molecules upon ligand binding.[10−18] Docking scoring functions are also being reparameterized to represent better water interactions in ligand−protein docking simulations.[19−21]

A different approach for the treatment of hydration in docking simulations is to attempt to model the role of individual ordered water molecules observed in X-ray crystal structures of proteins. Inclusion of these tightly bound water molecules in drug design methods has been investigated,[22−24] and various methods have been developed to include the interactions with water molecules in ligand−protein docking.[4,25,26] Barillari et al.[27] have used Monte Carlo replica exchange thermodynamic integration simulations to compute the absolute free energies of the binding of water molecules and to predict their displacement by a ligand.

Water molecules found in the crystal structures of proteins tend to occupy conserved positions in structurally related proteins[28−40] as well as in structures obtained under different conditions[41,42] and/or different bound ligands.[43−48] When a ligand binds to a protein, water molecules in the binding site either are displaced into the bulk solvent or are retained and mediate hydrogen-bonding interactions between ligand and protein, thereby stabilizing the ligand−protein complex.[27] A recent analysis of high-resolution crystal structures of ligand−protein complexes revealed that 85% of the complexes had one or more water molecules bridging the interactions between ligand and protein.[49] Consequently, ligand−protein docking simulations should seek to include, wherever possible, those water molecules observed in the binding site.

An important consideration is which water molecules need to be included for ligand−protein docking, particularly as any given water molecule that may play a critical role for the binding of one ligand may be otherwise displaced by other ligands. In such cases, the inclusion of a water molecule may lead to the incorrect prediction of docking poses. Recently, the effect of including water molecules in virtual screening was investigated, with a substantial increase in enrichment factors for 12 out of 24 targets reported, with the remaining targets remaining largely unaffected.[50]

Recently we reported a comprehensive survey of the effect of including water molecules in ligand−protein docking and found a statistically significant overall increase in accuracy when water molecules are included.[51] Here, we expand these studies by investigating the effect of including water molecules in cross-docking simulations, where a number of ligands taken from the crystal structures in a number of proteins have been docked in the presence of a common set of water molecules.

* Corresponding author. Phone: +61 892661017. Fax: +61 8 92662342. E-mail: R.Mancera@curtin.edu.au.

[§] Current address: Department of Chemistry, State University of New York at Binghamton, Binghamton, NY 13902.

10.1021/ci900345h
Published on Web 02/17/2010

METHODOLOGY

Selection of Ligand−Protein Complexes. Ligand-protein complexes were taken directly from the Protein Data Bank (PDB).[52] A number of proteins were initially selected on the basis that they had previously been determined to be cases where docking accuracy was found to improve significantly by the inclusion of water molecules.[51] The complexes were further selected on the basis that they contain water molecules in the binding site and that a large number of ligand−protein complexes have been crystallized. The proteins that were selected were penicillopepsin, lysozyme, fatty acid binding protein, D-xylose isomerase, renin and the SH2 domain of c-src tyrosine kinase.

Selection of Common Sets of Water Molecules. The crystal structure for each protein with the highest X-ray resolution was used as the template (reference), and all other crystal structures of the same protein were aligned to it by superposing all Cα atoms. Water molecule clusters present within a 5.0 Å radius with respect to the ligand in the template structure were visually inspected to select a common set of water molecules for each protein. Visual inspection was slightly extended beyond the cutoff radius to view the complete water clusters, especially when they were not fully seen at the edge of the cutoff radius (5.0 Å). Water molecules within 1.5 Å from the template water were considered to be in the same cluster. The orientation of the selected water molecules was optimized by energy minimization in the absence of any bound ligand, keeping all protein atoms rigid and fixing the position of the oxygen atom of each water molecule, as described below. Water molecules are identified by reference to their numbers in the PDB file of the reference structure.

Protein Structure Preparation. The ligand and any other cofactors in each protein structure were removed, and hydrogen atoms were added using Sybyl 8.0 (Tripos Inc.) with random orientations for water molecule hydrogens. The protonation states of all binding site residues were assigned, considering a pH of 7.0. In the case of penicillopepsin (an aspartyl protease), Asp213 was protonated. Only the selected water molecules (see above) were retained, and the remaining water molecules were removed. Optimization of hydrogen atoms was carried through energy minimization using the Tripos force field with a gradient convergence criterion of 0.001 kcal mol^{-1} $Å^{-1}$.

Ligand Preparation. Ligands were extracted from the superimposed complexes, and all bond and atom types were checked for consistency. Hydrogen atoms were added, and ionization states were determined, considering a pH of 7.0. In case of the phosphonate ligands of penicillopepsin, the unionized forms of the ligands were used.

Ligand−Protein Docking. In order to carry out a set of cross-docking simulations for each chosen protein, all ligands taken from the crystal structures of the same protein were docked onto the corresponding selected template structure. Docking simulations were performed using an implementation of the stochastic tunneling method,[53,54] using the PLP[55] scoring function to represent the free energy surface of the ligand−protein interaction. A bounding simulation box was defined around the binding site of each protein by considering a distance of 3.0 Å away from the outermost atoms of each bound ligand in each the *x*-, *y*- and *z*-directions. A total of 300 simulations of each ligand−protein complex were performed. All resulting binding poses were rescored using ScreenScore scoring function.[56]

Two sets of ligand−protein docking simulations were performed for each protein to determine the effect of the presence of common sets of water molecules on the accuracy of these cross-docking studies: one in the absence of water molecules and another one in the presence of the common set of water molecules. The predictive ability of the docking simulations was assessed by determining the root-mean-square deviation (rmsd) of the binding pose with the best energy of binding (the top-ranked solution). For consistency with the original Cambridge Crystallographic Data Centre (CCDC)/Astex data set[57] and with our own previous work,[51,54] the following qualitative classification of the accuracy of docking predictions was applied: (i) a *good* solution had a rmsd $\leq$ 1.0 Å; (ii) a *close* solution had a 1.0 < rmsd $\leq$ 2.0 Å; (ii) a solution *with errors* had a 2.0 < rmsd $\leq$ 3.0 Å; and (iv) a *wrong* solution had a rmsd > 3.0 Å. The best rmsd found during the docking simulations was used to determine the efficacy of the docking search method, as they represent the best-case scenario that could be achieved by the scoring function.

RESULTS AND DISCUSSION

Six different proteins were investigated, including two aspartyl proteases (penicillopepsin and human renin), lysozyme, fatty acid binding protein, D-xylose isomerase, and the SH2 domain of c-src tyrosine kinase. These proteins were selected as they had previously been determined to be cases where docking accuracy was found to improve significantly by the inclusion of water molecules.[51] These proteins, thus, represent cases where cross-docking of ligands from other crystal structures would allow us to determine if the judicious selection of a common set of water molecules would still result in improvements in the docking accuracy across a set of different ligands.

Penicillopepsin. Thirteen ligand−protein complexes with an X-ray resolution ranging from 0.95 to 1.80 Å are available for this protein. Water-mediated hydrogen-bonding interactions were observed between ligand and protein in all complexes (except 2web). All structures were aligned with respect to the template structure with the highest resolution (PDB code, 1bxo). Three water molecules (W511, W550, and W840) were observed in all of the complexes. The remaining seven water molecules were observed in most but not all of the complexes: W532 (12), W533 (10), W568 (10), W602 (9), W637 (12), W779 (10), and W845 (11), with the number of complexes containing each cluster of water molecules shown in the brackets. All ten water molecules were included in the docking simulations. W840 was involved in the water-mediated hydrogen-bonding interaction between ligand and protein.

Table 1 summarizes the results of all docking simulations for each ligand−protein complex. In the case of penicillopepsin, cross-docking in the absence of water molecules (NoW set) gave better results (as measured by an average rmsd $\leq$ 2.0 Å) for only two ligands: 1apw (correctly predicted in both cases) and 1bxq (the only ligand exhibiting significantly worse results when water molecules were considered). In the presence of water molecules, the scoring

Table 1. Results of Cross-Docking Simulations with and without Water Molecules[a]

	rmsd(Å)			rmsd (Å)	
PDB code	NoW	W	PDB code	NoW	W
penicillopepsin			D-xylose isomerase		
1apt	1.42	1.74	1xie	**5.01**	1.01
1apu	**6.27**	1.24	1xif	**2.95**	0.90
1apv	**8.27**	**2.12**	1xig	1.44	0.94
1apw	1.01	**2.01**	1xii	**2.12**	**4.86**
1bxo[b]	1.16	0.86	2glk[b]	**4.57**	1.90
1bxq	1.24	**6.01**	4xis[c]	1.12	0.82
1ppk	1.32	1.60		**3.66**	1.07
1ppl	**10.03**	1.45	renin		
1ppm	**11.05**	**5.00**	2g1o	0.77	0.57
2wea	**6.98**	**7.68**	2g1r	0.70	0.71
2web	**7.85**	**7.54**	2g1s	0.63	0.58
2wec	**9.42**	**4.57**	2g1y	0.76	0.89
2wed	**6.91**	**8.68**	2g21	0.59	0.56
lysozyme			2il2	1.00	**2.31**
1bb6	**7.39**	1.92	2i4q	0.87	0.87
1bb7	**4.83**	**10.32**	2iko	**8.74**	0.54
1lmc	**7.60**	1.46	2v0z	**3.79**	**3.27**
1lmo	**7.85**	**7.57**	1hrn[b]	**8.96**	**7.76**
1lmp	**9.30**	1.58	SH2 domain of c-src tyrosine kinase		
1lmq[b]	**11.54**	1.07	1o44	0.72	0.70
fatty acid binding protein			1o46	0.84	0.80
1icm[b]	2.00	**2.17**	1o47	**3.14**	0.82
1icn	**2.28**	**2.03**	1o48[b]	**3.40**	0.61
2ifb	**3.01**	1.83	1o49	0.77	1.00
			1o4b	**3.33**	1.04
			1a07	**4.38**	**2.85**
			1a1b	**3.84**	**5.24**
			1a1c	**2.62**	1.90
			1a1e	**2.82**	**4.27**

[a] Rmsd values >2 Å are shown in bold. [b] Template structure. [c] Structure 4xis had two ligands (xls393, xys394), both of which were docked separately.

function (ScreenScore)[56] predicted the wrong top-scoring binding pose (rmsd = 6.01 Å), with the polar amide groups in the ligand being orientated toward the water molecules. By contrast, the PLP[55] scoring function predicted the correct top-ranking binding pose (rmsd = 1.28 Å). The inclusion of water molecules largely improved the results for five ligands (1apu, 1apv, 1ppl, 1ppm, and 2wec), with rmsd values ≤2.1 Å (except for 1ppm and 2wec, which still gave wrong solutions). The inclusion of water molecules did not affect the docking accuracy to any significant extent for six ligands (correctly predicted: 1apt, 1bxo, and 1ppk; incorrectly predicted: 2wea, 2web, and 2wed). The poor docking results obtained for the ligands of 2wea, 2web, and 2wed, with or without water molecules, are due to steric clashes between the template protein side chains and ligand atoms. In the case of ligand 2wec, the poor docking results are due to steric clashes between the ligand and the water molecules W840 and W845. It is interesting to note that for ligand 1ppm, just as before, the ScreenScore scoring function predicted the wrong top-scoring binding pose (rmsd = 5.00 Å), whereas the PLP scoring function predicted the correct binding mode (rmsd = 1.25 Å).

Lysozyme. A total of six ligand−protein complexes with an X-ray resolution ranging from 1.60 to 2.00 Å are available for rainbow trout lysozyme (RBTL). Water-mediated hydrogen-bonding interactions were observed between ligand and protein in all complexes. All structures were aligned with respect to the template structure with the highest resolution

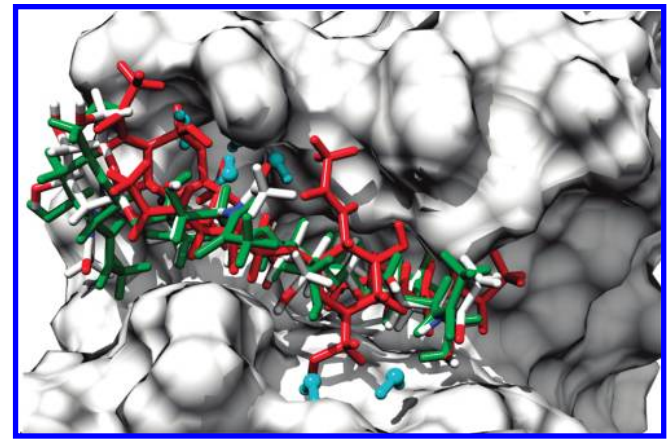

Figure 1. Docking results for the complex of lysozyme with the ligand in PDB structure 1lmq. The ligand in its native crystallographic binding mode is colored according to atom type. The top-ranked binding pose predicted in the presence of water molecules (W) is shown in green, and the top-ranked binding pose predicted in the absence of water molecules (NoW) is shown in red. Water molecules are shown in cyan.

(PDB code 1lmq). Four water molecules (W151, W152, W153, and W154) were observed in all of the complexes. Two water molecules (W150 and W155) were observed in five out of the six complexes. All six water molecules were included in the docking simulations.

The results of the docking simulations for each ligand−protein complex of lysozyme are summarized in Table 1. In the absence of water molecules, all six ligands were docked incorrectly. The inclusion of water molecules improved largely the docking accuracy of four ligands (1bb6, 1lmc, 1lmp, 1lmq), failing for ligands 1lmo and 1bb7 (in this case, the results were significantly worse). One of the hydroxyl groups in ligand 1lmo has a steric clash with the phenyl ring of Tyr62 in the protein, which explains the failure of the docking simulations to correctly predict its binding mode in the template structure. Figure 1 illustrates and compares the crystallographic binding mode, the top-ranked solution predicted in the presence of water molecules (W), and the top-ranked solution predicted in the absence of water molecules (NoW) for the ligand in PDB structure 1lmq. The presence of water molecules sterically prevents the placement of several functional groups of this ligand in positions otherwise favored in simulations without water molecules (rmsd = 11.54 Å), resulting in a correct binding mode (rmsd = 1.07 Å).

Fatty Acid Binding Protein. Three ligand−protein complexes with an X-ray resolution ranging from 1.50 to 2.00 Å are available for rat intestinal fatty acid binding protein. Water-mediated hydrogen-bonding interactions were observed between ligand and protein in all complexes (except 1icn). All structures were aligned with respect to the template structure with the highest resolution (PDB code 1icm). Five water molecules (W139, W167, W168, W183, and W193) were observed in all of the complexes, with an additional water molecule (W170) observed in PDB structure 2ifb but not in 1icn complex. All six water molecules were included in the docking simulations.

Table 1 contains a summary of the results of the docking simulations for each ligand−protein complex of this fatty acid binding protein. In the absence of water molecules, two ligands (1icn and 2ifb) were predicted incorrectly (rmsd ≥

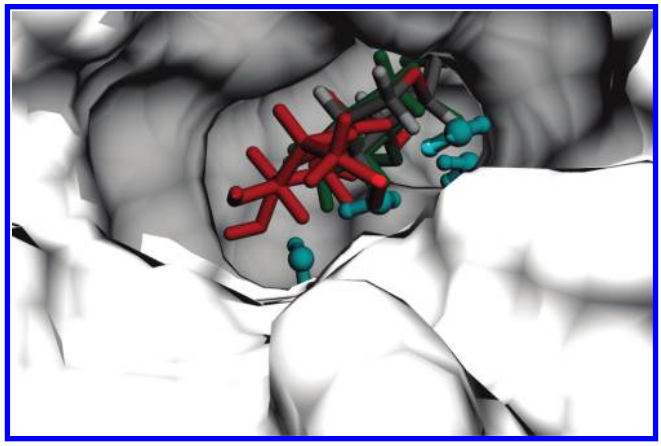

Figure 2. Docking results for the complex of D-xylose isomerase with the ligand in PDB structure 1xie. The ligand in its native crystallographic binding mode is colored according to atom type. The top-ranked binding pose predicted in the presence of water molecules (W) is shown in green, and the top-ranked binding pose predicted in the absence of water molecules (NoW) is shown in red. Water molecules are shown in cyan.

2 Å), and only one ligand (1icm) was correctly predicted (rmsd = 2.00 Å). The addition of water molecules resulted in a significant improvement in the docking accuracy for ligand 2ifb (rmsd improved from 3.01 to 1.83 Å). The best binding poses for the other two ligands were very similar to those predicted by docking in the absence of water molecules and were only slightly wrong (rmsd of 2.17 Å for 1icm and 2.03 Å for 1icn).

D-Xylose Isomerase. There are 11 ligand−protein complexes with an X-ray resolution ranging from 0.94 to 1.90 Å available for D-xylose isomerase of *Streptomyces rubiginosus*. The eight complexes that contain two metal atoms (Mn) in the active site were chosen.[58] All of the structures were aligned with respect to the template structure with the highest resolution (PDB code 2glk). Four water molecules (W1048, W1119, W1271, and W1316) were found in all eight complexes. Three water molecules (W1028, W1365, and W1387) were found in seven complexes. All seven water molecules were included in the docking simulations. Ligand 9xia is a covalent inactivator[43] that is bound to His54. Ligand 1xid was found to clash with water molecules W1387 and W1365, which were part of the common set of water molecules defined for this protein. Consequently, these two ligands were not considered in the docking simulations. The six remaining ligand−protein complexes have an X-ray resolution ranging from 0.94 to 1.7 Å. Water-mediated hydrogen-bonding interactions were observed between ligand and protein in all complexes. The two metal ions (Mn) were retained during the simulations. The protonated form of His54 was considered for the docking simulations, as recent neutron diffraction studies revealed that this residue is doubly protonated.[59] PDB structure 4xis has two ligands (xls393 and xys394), which were docked separately.

The results of the docking simulations for each ligand−protein complex of D-xylose isomerase are summarized in Table 1. In the absence of water molecules, only three ligands (1xig, 1xii, and 4xis (xls393)) were predicted correctly (rmsd ≤ 2.1 Å). Six out of seven ligands were predicted correctly in the presence of water molecules, with only ligand 1xii showing a worse prediction. Figure 2 illustrates and compares the crystallographic binding mode, the top-ranked solution predicted in the presence of water molecules (W), and the top-ranked solution predicted in the absence of water molecules (NoW) for the ligand in PDB structure 1xie. The errors in the binding mode of this ligand arise due to the alternative orientation of the sugar ring of the ligand. The top-ranked binding pose obtained in the absence of water molecules placed the ligand in an orientation where it lacks interactions with key residues of the protein. When the docking simulations were performed in the presence of water molecules, the native binding mode was reproduced fairly closely.

Human Renin. Many ligand−protein complexes are available for this protein, and so only ten complexes with an X-ray resolution ranging from 1.80 Å to 2.70 Å were randomly chosen. Water-mediated hydrogen-bonding interactions were observed between ligand and protein in all complexes (except 2g1o). All of the structures were aligned with respect to the template structure with the highest resolution (PDB code 1hrn, chain A). Water molecule W866 was found in all complexes. Water molecule W118 (ID with respect to PDB structure 2iko) was observed in eight complexes but was missing in the template structure and in the PDB structure 2v0z. An oxygen atom in the ligand in the template structure displaces this water molecule. In the case of PDB structure 2v0z, W118 would have a steric clash with a carbon atom in the ligand. Water molecule W25 (ID with respect to PDB structure 2g21) was observed in eight complexes but was displaced by a SO_2 group in ligand 2il2. W25 also would have a steric clash with ligand 1hrn. W118 was merged with the template structure before optimization of the orientation of the water molecules. It was determined that W866 and W118 (PDB code 2iko) are within hydrogen-bonding distance of several ligands and were, thus, included in the docking simulations. However, W866 was only used for the docking of 1hrn and 2v0z ligands.

A summary of the results of the docking simulations for each ligand−protein complex of human renin can be found in Table 1. In the absence of water molecules, seven out of ten ligands were predicted correctly (rmsd ≤ 2.0 Å), with the remaining three ligands (2iko, 2v0z, and 1hrn) being incorrectly predicted. The inclusion of water molecules allowed the correct prediction of ligand 2iko but resulted in no significant improvement for ligands 2v0z and 1hrn. Ligand 2il2 was the only ligand with which the inclusion of water molecules resulted in a (slight) worsening of the docking predictions (rmsd = 1.00 and 2.31 Å for NoW and W, respectively), with the orientation of its naphthyl sulphonyl amino moiety deviating from its native conformation.

SH2 Domain of c-Src Tyrosine Kinase. A large number of ligand−protein complexes are available for this protein and consequently only ten complexes with an X-ray resolution ranging from 1.55 to 2.40 Å were randomly chosen. Water-mediated hydrogen-bonding interactions were observed between ligand and protein in all complexes. All structures were aligned with respect to the template structure with the highest resolution (PDB code 1o48). Three water molecules (W3, W99, and W110) were observed in all of the complexes. Water molecule W21 was absent in PDB structure 1a1b, and water molecule W92 was absent in PDB structures 1a1c and 1a07. In addition, water molecule W180 was found in six complexes but was missing in PDB

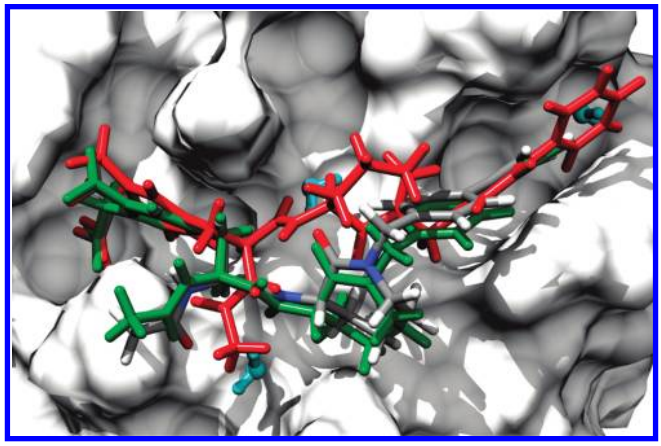

Figure 3. Docking results for the complex of the SH2 domain of c-src tyrosine kinase with ligand 1o48. The ligand in its native crystallographic binding mode is colored according to atom type. The top-ranked binding pose predicted in the presence of water molecules (W) is shown in green, and the top-ranked binding pose predicted in the absence of water molecules (NoW) is shown in red. Water molecules are shown in cyan.

structures 1a1e, 1a1b, 1a1c, and 1a07, with the carboxylate group of Glu102 of these peptide ligands occupying this position. Water molecules W3, W21, W99, and W110 were included in the docking simulations.

Table 1 summarizes the results of the docking simulations for each ligand−protein complex of this protein. In the absence of water molecules, only three ligands (1o44, 1o46, and 1o49) were predicted correctly, with no significant changes upon inclusion of the water molecules. The addition of water molecules largely improved the docking predictions for four ligands (1o47, 1o48, 1o4b, and 1a1c. Three ligands (1a07, 1a1b, and 1a1e) were found to be docked incorrectly in both the presence and the absence of water molecules. Figure 3 illustrates and compares the crystallographic binding mode, the top-ranked solution predicted in the presence of water molecules (W), and the top-ranked solution predicted in the absence of water molecules (NoW) for the ligand in PDB structure 1o48. The errors in the predicted binding mode of this ligand in the absence of water molecules arise due to the alternative orientation of the phenyl ring and the linker group in the ligand. When the docking simulations were performed in the presence of water molecules, the native binding mode was reproduced fairly closely.

Overall Results. Table 2 summarizes the efficacy of the docking search method in the form of the average best rmsd obtained for each protein, clearly revealing that in both the absence and the presence of water molecules, the docking search method was highly successful in finding binding poses close to the crystallographic binding modes (the highest mean rmsd is 1.32 Å for docking simulations of the fatty acid binding protein in the presence of water molecules), with only a marginal improvement of the overall average best rmsd from 1.05 to 1.00 Å. Table 2 also summarizes the differences in the accuracy of the cross-docking simulations in the form of the average rmsd for the top-ranked binding poses predicted for each protein. The overall average rmsd of the top-ranked binding poses improved in the presence of water molecules from 4.13 to 2.64 Å, with the difference shown to be statistically significant with a paired *t*-test (*p*-value = 0.001355). These results show that the accuracy improved significantly in the presence of water molecules

Table 2. Mean Best Rmsd and Mean Rmsd of Top Energy-Ranked Solutions Obtained for Each Protein in the Absence and the Presence of Water Molecules

	mean best rmsd (Å)		mean rmsd of top-ranked solution (Å)	
protein	NoW	W	NoW	W
penicillopepsin	1.21	1.28	5.61	3.88
lysozyme	1.12	0.90	8.08	3.99
fatty acid binding protein	1.31	1.32	2.43	2.01
D-xylose isomerase	0.91	0.65	2.97	1.64
renin	0.70	0.71	2.68	1.81
SH2 domain of c-src tyrosine kinase	1.18	1.13	2.58	1.92
average	1.05	1.00	4.13	2.64
p-value			0.001355	

Table 3. Number of Ligands Docked Successfully (Rmsd ≤ 2.2 Å) in Cross-Docking Simulations

		number of ligands docked successfully	
target	total number of ligands	NoW	W
penicillopepsin	13	5	7
lysozyme	6	0	4
fatty acid binding protein	3	1	3
D-xylose isomerase	7	3	6
renin	10	7	7
SH2 domain of c-src tyrosine kinase	10	3	7
total number	49	19	34
Success rate (%)		38.77	69.39

for all six targets, with the improvement being more noticeable for penicillopepsin, lysozyme, and D-xylose isomerase.

Table 3 summarizes all cross-docking simulations by reporting the number of ligands that were successfully docked with each protein. Successful docking simulations were considered to be those whose top-ranked binding pose was within a 2.2 Å rmsd of the crystallographic binding mode. A success rate of 39% (19 ligands out of 49) was achieved in the absence of water molecules (NoW), increasing substantially to 69% (34 ligands out of 49) in the presence of water molecules (W).

CONCLUSIONS

Cross-docking simulations on six different proteins were found to show a significant improvement in the accuracy of predicted binding modes when conserved water molecules in the binding site were included, as revealed by statistically significant reductions in the mean rmsd of the top-ranked binding poses with respect to the crystallographic binding modes. Penicillopepsin, lysozyme, and D-xylose isomerase showed the most marked improvements in docking accuracy upon inclusion of water molecules, whereas human rennin, the SH2 domain of c-src tyrosine kinase, and the fatty acid binding protein show smaller improvements in docking accuracy in the presence of water molecules.

The X-ray structures of lysozyme and D-xylose isomerase contain carbohydrate-based ligands, whose functional groups were involved in water-mediated hydrogen-bonding interac-

tions with the protein. Consequently, the inclusion of water molecules in the binding site had a significant positive impact on docking accuracy. The structures of the SH2 domain of c-src tyrosine kinase with its ligands revealed that the ligands also have water-mediated hydrogen-bonding interactions with the protein, although only ligands similar (except 1a1c) in structure to the template ligand were docked successfully.

Poor cross-docking results were obtained for a number of ligands in penicillopepsin, where the inclusion of water molecules did not improve largely the accuracy of the predictions. This was due mainly to the presence of steric clashes with both the amino acid side chains in the binding site and the water molecules of the template molecule.

The vast majority of renin ligands were predicted accurately in the cross-docking simulations in the absence of water molecules, and so the inclusion of water molecules is not warranted for this protein. In the case of the fatty acid binding protein, the inclusion of water molecules appeared to have a beneficial effect on cross-docking accuracy; however, the small number of ligands investigated limits the significance of these findings.

The inclusion of water molecules in ligand−protein docking requires an initial judicious selection of common sets of water molecules across different ligand−protein complexes. This, in turn, is likely to result in significant increases in docking accuracy in the absence of conformational changes in the binding site, such as the occurrence of different rotameric states for amino acid side chains.

ACKNOWLEDGMENT

We thank the Interactive Virtual Environments Centre (iVEC) of Western Australia for the award of an Industry and Government Uptake Grant.

Supporting Information Available: Results of cross docking simulations for each individual ligand−protein complex with and without water molecules.This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. D. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
(2) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377–395.
(3) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209–225.
(4) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
(5) Wasserman, Z. R.; Hodge, C. N. Fitting an inhibitor into the active site of thermolysin: a molecular dynamics case study. *Proteins* **1996**, *24*, 227–237.
(6) Di Nola, A.; Roccatano, D.; Berendsen, H. J. C. Molecular dynamics simulation of the docking of substrates to proteins. *Proteins* **1994**, *19*, 174–182.
(7) Mangoni, R.; Roccatano, D.; Di Nola, A. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins* **1999**, *35*, 153–162.
(8) Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Sidechain flexibility in proteins upon ligand binding. *Proteins* **2000**, *39*, 261–268.
(9) Mancera, R. L. Molecular modeling of hydration in drug design. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 275–280.
(10) Li, Z.; Lazaridis, T. Thermodynamic contributions of the ordered water molecule in HIV-1 protease. *J. Am. Chem. Soc.* **2003**, *125*, 6636–6637.
(11) Fornabaio, M.; Spyrakis, F.; Mozzarelli, A.; Cozzini, P.; Abraham, D. J.; Kellogg, G. E. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. *J. Med. Chem.* **2004**, *47*, 4507–4516.
(12) Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Free energy of ligand binding to protein: evaluation of the contribution of water molecules by computational methods. *Curr. Med. Chem.* **2004**, *11*, 3093–3118.
(13) Lu, Y.; Yang, C.-Y.; Wang, S. Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes. *J. Am. Chem. Soc.* **2006**, *128*, 11830–11839.
(14) Helms, V.; Wade, R. C. Thermodynamics of water mediating protein-ligand interactions in cytochrome P450CAM. A molecular dynamics study. *Biophys. J.* **1995**, *69*, 810–824.
(15) Helms, V.; Wade, R. C. Computational alchemy to calculate absolute protein-ligand binding free energy. *J. Am. Chem. Soc.* **1998**, *120*, 2710–2713.
(16) Hamelberg, D.; McCammon, J. A. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.
(17) Li, Z.; Lazaridis, T. The effect of water displacement on binding thermodynamics: concanavalin A. *J. Phys. Chem. B* **2005**, *109*, 662–670.
(18) Li, Z.; Lazaridis, T. Thermodynamics of buried water clusters at a protein-ligand interface. *J. Phys. Chem. B* **2006**, *110*, 1464–1475.
(19) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902.
(20) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision Glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.
(21) Mancera, R. L. A new explicit hydration penalty score for ligand-protein interactions. *Chem. Phys. Lett.* **2004**, *399*, 271–275.
(22) Mancera, R. L. De novo ligand design with explicit water molecules: an application to bacterial neuraminidase. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 479–499.
(23) Lloyd, D. G.; García-Sosa, A. T.; Alberts, I. L.; Todorov, N. P.; Mancera, R. L. The effect of tightly-bound water molecules on the structural interpretation of ligand-derived pharmacophore models. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 89–100.
(24) García-Sosa, A. T.; Mancera, R. L. The effect of tightly-bound water molecules on scaffold diversity in computer-aided de novo ligand design of CDK-2 inhibitors. *J. Mol. Model.* **2006**, *12*, 422–431.
(25) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins* **1999**, *34*, 17–28.
(26) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002**, *46*, 34–40.
(27) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.* **2007**, *129*, 2577–2587.
(28) Chung, E.; Henriques, D.; Renzoni, D.; Zvelebil, M.; Bradshaw, J. M.; Waksman, G.; Robinson, C. V.; Ladbury, J. E. Mass spectrometric and thermodynamic studies reveal the role of water molecules in complexes formed between SH2 domains and tyrosyl phosphopeptides. *Struct. Fold. Des.* **1998**, *6*, 1141–1151.
(29) Baker, E. N.; Hubbard, R. E. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97–179.
(30) Sreenivasan, U.; Axelsen, P. H. Buried water in homologous serine proteases. *Biochemistry* **1992**, *31*, 12785–12791.
(31) Loris, R.; Stas, P. P.; Wyns, L. Conserved waters in legume lectin crystal structures. The importance of bound water for the sequence structure relationship within the legume lectin family. *J. Biol. Chem.* **1994**, *269*, 26722–26733.
(32) Shaltiel, S.; Cox, S.; Taylor, S. S. Conserved water molecules contribute to the extensive network of interactions at the active site of protein kinase A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 484–491.
(33) Sanschagrin, P. C.; Kuhn, L. A. Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. *Protein Sci.* **1998**, *7*, 2054–2064.
(34) Krem, M. M.; Di Cera, E. Conserved water molecules in the specificity pocket of serine proteases and the molecular mechanism of Na+ binding. *Proteins* **1998**, *30*, 34–42.
(35) Loris, R.; Langhorst, U.; De Vos, S.; Decanniere, K.; Bouckaert, J.; Maes, D.; Transue, T. R.; Steyaert, J. Conserved water molecules in

a large family of microbial ribonucleases. *Proteins* **1999**, *36*, 117–134.
(36) Ogata, K.; Wodak, S. J. Conserved water molecules in MHC class-I molecules and their putative structural and functional roles. *Protein Eng.* **2002**, *15*, 697–705.
(37) Bottoms, C. A.; Smith, P. E.; Tanner, J. J. A structurally conserved water molecule in Rossmann dinucleotide-binding domains. *Protein Sci.* **2002**, *11*, 2125–2137.
(38) Prasad, B. V. L. S.; Suguna, K. Role of water molecules in the structure and function of aspartic proteinases. *Acta Crystallogr., Sect. D: Biol.Crystallogr.* **2002**, *58*, 250–259.
(39) Bottoms, C. A.; Schuermann, J. P.; Agah, S.; Henzl, M. T.; Tanner, J. J. Crystal structure of rat R-parvalbumin at 1.05 Å resolution. *Protein Sci.* **2004**, *13*, 1724–1734.
(40) Bottoms, C. A.; White, T. A.; Tanner, J. J. Exploring structurally conserved solvent sites in protein families. *Proteins* **2006**, *64*, 404–421.
(41) Nakasako, M. Large-scale networks of hydration water molecules around bovine beta-trypsin revealed by cryogenic X-ray crystal structure analysis. *J. Mol. Biol.* **1999**, *289*, 547–564.
(42) Zhang, X.-J.; Matthews, B. W. Conservation of solvent-binding sites in 10 crystal forms of T4 lysozyme. *Protein Sci.* **1994**, *3*, 1031–1039.
(43) Carrell, H. L.; Glusker, J. P.; Burger, V.; Manfre, F.; Tritsch, D.; Biellmann, J.-F. X-ray analysis of D-xylose isomerase at 1.9 Å: native enzyme in complex with substrate and with a mechanism-designed inactivator. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 4440–4444.
(44) Faerman, C. H.; Karplus, P. A. Consensus preferred hydration sites in six FKBP12-drug complexes. *Proteins* **1995**, *23*, 1–11.
(45) Poornima, C. S.; Dean, P. M. Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 521–531.
(46) Babor, M.; Sobolev, V.; Edelman, M. Conserved positions for ribose recognition: importance of water bridging interactions among ATP, ADP and FAD-protein complexes. *J. Mol. Biol.* **2002**, *323*, 523–532.
(47) Powers, R. A.; Shoichet, B. K. Structure-based approach for binding site identification on AmpC beta-lactamase. *J. Med. Chem.* **2002**, *45*, 3222–3234.
(48) Mustata, G.; Briggs, J. M. Cluster analysis of water molecules in alanine racemase and their putative structural role. *Protein Eng.* **2004**, *17*, 223–234.
(49) Lu, Y.; Wang, R.; Yang, C. Y.; Wang, S. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J. Chem. Inf. Model.* **2007**, *47*, 668–675.
(50) Huang, N.; Shoichet, B. K. Exploiting ordered waters in molecular docking. *J. Med. Chem.* **2008**, *51*, 4862–4865.
(51) Roberts, B. C.; Mancera, R. L. Ligand-protein docking with water molecules. *J. Chem. Inf. Model.* **2008**, *48*, 397–408.
(52) Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the worldwide protein data bank. *Nat. Struct. Biol.* **2003**, *10*, 980.
(53) Todorov, N. P.; Mancera, R. L.; Monthoux, P. H. A new quantum stochastic tunneling method for ligand protein docking. *Chem. Phys. Lett.* **2003**, *369*, 257–263.
(54) Mancera, R. L.; Källblad, P. K.; Todorov, N. P. Ligand-protein docking using a quantum stochastic tunneling optimization method. *J. Comput. Chem.* **2004**, *25*, 858–864.
(55) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317–324.
(56) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
(57) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein ligand interaction. *Proteins* **2002**, *49*, 457–471.
(58) A number of identical ligand−protein complexes were removed (PDB codes 3xis, 2xis, 1xic, and 8xia), leaving eleven complexes with different ligands. Among them, two ligand−protein complexes (PDB code 1xih and 1xij) contain only one metal (Mn) ion in the active site. In addition, PDB structure 1gw9 contains two calcium ions in the active site. Thus, only eight ligand−protein complexes containing two metal (Mn) ions were considered.
(59) Katz, A. K.; Li, X.; Carrell, H. L.; Hanson, B. L.; Langan, P.; Coates, L.; Schoenborn, B. P.; Glusker, J. P.; Bunick, G. J. Locating active-site hydrogen atoms in D-xylose isomerase: Time-of-flight neutron diffraction. *Proc. Nat. Acad. Sci. U.S.A.* **2006**, *103*, 8342–8347.

CI900345H