BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 19 2006, pages 2340–2347
doi:10.1093/bioinformatics/btl395

Structural bioinformatics

Solvated docking: introducing water into the modelling of biomolecular complexes

Aalt D. J. van Dijk and Alexandre M. J. J. Bonvin*

Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, 3584CH, Utrecht,
The Netherlands

Received on May 10, 2005; revised on June 26, 2006; accepted on July 17, 2006
Advance Access publication August 9, 2006
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Interfacial water, which plays an important role in mediating biomolecular interactions, has been neglected in the modelling of biomolecular complexes.

Methods: We present a solvated docking approach that explicitly accounts for the presence of water in protein–protein complexes. Our solvated docking protocol is based on the concept of the first encounter complex in which a water layer is present in-between the molecules. It mimics the pathway from this initial complex towards the final assembly in which most waters have been expelled from the interface. Docking is performed from solvated biomolecules and waters are removed in a biased Monte Carlo procedure based on water-mediated contact propensities obtained from an analysis of high-resolution crystal structures.

Results: We demonstrate the feasibility of this approach for protein–protein complexes representing both 'wet' and 'dry' interfaces. Solvated docking leads to improvements both in quality and scoring. Water molecules are recovered that closely match the ones in the crystal structures.

Availabilty: Solvated docking will be made available in the future release of HADDOCK version 2.0 (http://www.nmr.chem.uu.nl/haddock).

Contact: a.m.j.j.bonvin@chem.uu.nl

Supplementary information: Supplementary Data are available at *Bioinformatics* Online.

1 INTRODUCTION

The modelling of protein–protein complexes by means of docking (a computational approach which models the unknown structure of a complex from its constituents) has become increasingly popular, as witnessed by the CAPRI (Critical Assessment of PRedicted Interactions) experiment (Mendez *et al*., 2005). Docking approaches have benefited from knowledge obtained by detailed analyses of binding interfaces (Halperin *et al*., 2002; van Dijk *et al*., 2005a). As discussed in a recent review, water molecules are expected to influence the assembly of biomolecular complexes (Chandler, 2005), and, as such, to be important for protein–protein docking. An analysis based on Voronoi volume showed that only upon inclusion of interfacial solvent molecules are protein–protein interfaces as densely packed as protein interiors (Lo Conte *et al*., 1999). So far, however, water has been neglected generally in biomolecular docking. Its role and importance in single proteins have been discussed (Rashin *et al*., 1986; Wade *et al*., 1993; Wade and Goodford, 1993; Hubbard *et al*., 1994; Robert and Ho, 1995; Raschke, 2006) and several case studies have analysed its conservation in 3D structures of homologues (Sreenivasan and Axelsen, 1992; Zhang and Matthews, 1994; Robert and Ho, 1995; Tame *et al*., 1996; Carugo, 1999; Carugo and Bordo, 1999; Loris *et al*., 1999; Babor *et al*., 2002; Houborg *et al*., 2003; Mustata and Briggs, 2004). There has also been quite someinterest in identifying and predicting the positions of water molecules in known structures: this can be quite successfully performed, for example, by GRID (Boobbyer *et al*., 1989; Wade *et al*., 1993; Wade and Goodford, 1993) or Fold-X (Schymkowitz *et al*., 2005). These kind of approaches, however, are not very well suited for docking purposes, since the structure of the complex is not known a priori. Ideally, water should be accounted for directly during the docking process since its presence might affect the resulting models. So far this has only be done for protein–ligand (Rejto and Verkhivker, 1997; Rarey *et al*., 1999; Osterberg *et al*., 2002; Yang and Chen, 2004; Verdonk *et al*., 2005) and nucleic acid–ligand docking (Moitessier *et al*., 2006) .

Only very recently has the role of water molecules at protein–protein interfaces been investigated. A hydrogen bonding potential for water-mediated contacts, in combination with a solvated rotamer library for describing side chain conformations, has been shown to predict rather successfully the positions of water molecules in complexes with known structures (Jiang *et al*., 2005). In another study (Rodier *et al*., 2005), various properties of interfacial water molecules such as residue preference and their number per unit of interface area were investigated.

We have experimented previously with the inclusion of water in the NMR structure calculation of a protein–non-specific DNA complex (Kalodimos *et al*., 2004): in that case, an extensive set of NOEs could be used, which forced the solvated biomolecules to come together and the unnecessary waters to leave the interface in a simulated annealing molecular dynamic approach. In general, in docking, this kind of experimental information is not available and, in the absence of a driving force, the water molecules will remain trapped at the interface. Alternative approaches are thus needed to remove the unnecessary water molecules from the interface. We have developed for this purpose a solvated docking protocol implemented in our data-driven docking approach HADDOCK (Dominguez *et al*., 2003) and demonstrate here for the first time that water can be explicitly included in protein–protein docking.

*To whom correspondence should be addressed.

2 METHODS

2.1 Database analysis

In order to obtain information on water in high-resolution crystal structures of complexes, the non-redundant dataset of Keskin *et al*. (2004) was analysed using CNS (Brunger *et al*., 1998) and a set of home written Python scripts. Interface residues were defined as residues having at least one heavy-atom contact with a residue from the partner chain, within a 10 Å cut-off distance. Water-mediated contacts were defined between pairs of interface residues, provided a water molecule is making at least one heavy-atom contact within 5 Å with both residues. Water-mediated contacts were designated main chain when at least one contact was made via a backbone atom; otherwise they were designated side chain.

To investigate whether the various types of water-mediated contacts adopt specific, well-defined conformations, we clustered them on the basis of positional RMSD values: the RMSD values were calculated after least-square positional fitting on the coordinates of the water oxygen, its contacting heavy atoms within 5 Å on both chains and their respective first bonded partner (total of five atoms). Since several atoms of a given side-chain can make contacts with the water oxygen atom within 5 Å, various combinations of atoms were tested for the calculation of the RMSD matrix and the one resulting in the best clustering (most populated first cluster) was selected for each amino acid–amino acid pair. Clustering was performed separately for main chain–water–main chain, side chain–water–side chain and main chain–water–side chain contacts. In the case of main chain contacts, N and O were defined as contacting atoms, with CA and C, respectively, as bonded neighbours.

RMSDs were calculated using g_rms (Lindahl *et al*., 2001) and Profit (www.bioinf.org.uk/software/profit). Clustering was performed using the greedy algorithm described by Daura *et al*. (1999), with a cut-off of 1.5 Å. This cut-off was based on an analysis of the distribution of all RMSD values (data not shown). Contacts involving two close waters thatwould fall into the same cluster were counted only once.

2.2 Protein–protein docking using explicit water

HADDOCK incorporates information about the interface in ambiguous interaction restraints (AIRs) that drive the docking. An AIR is defined as an ambiguous intermolecular distance (d_{iAB}) with a maximum value of typically 2 Å between any atom m of an active residue i of protein A (m_{iA}) and any atom n of both active and passive residues k (N_{res} in total) of protein B (n_{kB}) (and inversely for protein B). The effective distance d_{iAB}^{eff} for each restraint is calculated using the following equation:

$$d_{iAB}^{\mathrm{eff}} = \left(\sum_{m_{iA}=1}^{N_{\mathrm{atoms}}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{\mathrm{atoms}}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-\frac{1}{6}},$$

where N_{atoms} indicates all atoms of a given residue and N_{res} the sum of active and passive residues for a given molecule. Note that the effective distance calculated in this way will always be shorter than the shortest distance entering the sum, which is the reason why we can use a rather short upper bound of 2 Å. The definition of passive residues ensures that residues which are at the interface but are not detected are still able to satisfy the AIR restraints, i.e. contact active residues of the partner molecule. For details see Dominguez *et al*. (2003) and van Dijk *et al*. (2005a). HADDOCK consists of a collection of scripts derived from ARIA1.2 (Linge *et al*., 2003a) and CNS (Brunger *et al*., 1998). The respective position and orientation of the two molecules are first randomized. Then docking is performed consisting of a rigid body energy minimization, followed by semi-flexible simulated annealing in torsion angle space and final refinement in explicit solvent. Rigid body docking is performed a number of times (1000); each time, out of a number of trials (typically 5) only the best model is selected and written to disk.

We modified the rigid body docking stage to explicitly include water. We start by solvating the two chains using a box of TIP3P (Jorgensen *et al*., 1983) water. All waters outside a cut-off range (<4.0 Å to >8.0 Å) from the protein are removed. A short molecular dynamics (MD) run is performed to optimize the water positions while keeping the proteins fixed (4000 MD steps consisting of four times 1000 steps at a temperature of 600, 500, 400 and 300 K, respectively). After that, all waters further away than 5.5 Å are removed. An ensemble of different solvation shells (typically 5) is generated by randomly rotating the protein before adding the solvation shell. We also experimented with the use of GRID (Boobbyer *et al*., 1989) to place the initial waters around the separate protein chains. The results of the subsequent docking did not depend much on the choice of the solvating method (data not shown). The solvated docking protocol itself is presented in the Results section.

The standard semi-flexible refinement of HADDOCK consists of two rigid body simulated annealing stages followed by two simulated annealing stages with flexibility introduced first on side chains and then on backbone. For solvated docking we only used the latter two semi-flexible simulated annealing stages.

Non-bonded energies (sum of van der Waals and electrostatic terms) are calculated with an 8.5 Å distance cut-off using the OPLS non-bonded parameters (Jorgensen and Tirado-rives, 1988) from the parallhdg5.3.pro parameter file (Linge *et al*., 2003b); the dielectric constant ε is set to 10.0 to damp the electrostatic contribution in vacuum. The overall score is calculated as a weighted sum of different terms, using the default HADDOCK2.0 values for the weights (rigid body stage: E_{vdW} 0.01, E_{elec} 1.0, E_{AIR} 0.01, BSA −0.01, E_{desolv} 1.0; semi-flexible refinement: E_{vdW} 1.0, E_{elec} 1.0, E_{AIR} 0.1, BSA −0.01, E_{desolv} 1.0). Here vdW is van der Waals energy; elec, electrostatic energy; AIR, ambiguous interaction restraints; BSA, buried surface area; and desolv, desolvation energy. The desolvation energy is calculated using the atomic desolvation parameters of Fernandez-Recio *et al*. (2004). The various weights were obtained by a grid search to optimize scoring over the complexes tested so far including CAPRI targets. These were optimized separately for the various stages of HADDOCK to reflect the various levels of complexity and refinement (from rigid body docking in vacuum to flexible refinement in explicit solvent).

2.3 Test systems

We tested our protocol on 10 protein–protein complexes (Table 2). Note that there are only a limited number of complexes that are suitable as test cases: the resolution should be high enough (>2 Å) in order to have reliable positions for interfacial water molecules, and the free structures of the components of the complex should be available. We used all structures from the docking benchmark (Mintseris *et al*., 2005) satisfying those criteria and a few other complexes which we have been testing before. For two of these, E2A–HPr (Wang *et al*., 2000) and cohesin–dockerin (Carvalho *et al*., 2003), we used experimental data available from the literature (NMR chemical shift perturbation data for E2A–HPr (Dominguez *et al*., 2003) and mutagenesis and conservation data as used previously for docking cohesin–dockerin, which was one of the targets in round 4 of CAPRI (van Dijk *et al*., 2005b). For the others, AIRs were defined based on the interface residues identified in the crystal structure; for those complexes, to simulate a more realistic case, 50% of the restraints were randomly removed for each docking trial. When free structures of the complex components were available (seven cases, Table 2), we performed unbound docking followed by semi-flexible refinement as well as bound docking. For cohesin–dockerin, bound–unbound docking was performed in addition to bound docking, and for the other two cases only bound docking was performed.

3 RESULTS

Our ‘solvated docking’ protocol is based on the physical concept that, in the first encounter complex, a water layer will be present in-between the two protein chains. To proceed from the encounter complex to the final structure, most of the interfacial waters have to be removed. Our protocol mimics this process by starting the docking from solvated molecules. Water is subsequently

Table 1. Analysis of water in non-redundant Keskin dataset

Resolution (Å)	1.1–2.0
$N_{\text{structures}}$	19
<Chain length>	158 (111)
<No. of waters>	346 (264)
<No. of water per residue>	0.80
<Number of water at interface>[a]	24 (18)
N_c/N_{wmc}^b	7155/1544
f_{wmc}: sc/mc[c]	0.16/0.05

[a] Number of waters within heavy atom distance cut-off of 5.0 Å from both chains.

[b] N_c, total number of interface contacts (defined using a 10 Å heavy atom distance cut-off) in dataset; N_{wmc}, total number of water-mediated contacts in dataset.

[c] f_{wmc}, fraction of water-mediated side chain (sc) and main chain (mc) contacts.

removed in a biased Monte Carlo procedure based on water-mediated contact propensities. The latter are obtained from an analysis of a database of high-resolution crystal structures of protein–protein complexes. In the following we will first describe the results of this analysis and then present our solvated docking protocol, demonstrating its feasibility for a number of protein–protein complexes.

3.1 Analysis of water mediated contacts

In order to extract statistics of water-mediated contacts, we analysed the high-resolution structures ($\leq$2.0 Å) in the non-redundant dataset of protein–protein interfaces of Keskin *et al*. (2004). The corresponding PDB id's are provided in Supplementary Table 5. Some general statistics of our dataset are listed in Table 1.

In Figure 1, the fraction of water-mediated side chain and main chain contacts for all 20 × 20 amino acid combinations is shown. It is clear from this figure that preferences do exist for specific water-mediated contacts, an information which should be useful in the modelling of protein–protein complexes by docking (see below). In order to assess the statistical significance of the fractions of water-mediated contacts we compared the values obtained from the non-redundant filtered set with those obtained using the complete redundant set of structural homologues. Since these have a lower resolution, the derived fractions are lower than those from the filtered set (data not shown); there is, however, a clear correlation between the two datasets ($R = 0.6$). It is, however, clear that the propensities reported here should be refined in the future by making use of the (rather slowly) increasing number of protein complexes deposited into the PDB.

To find out whether interfacial water molecules adopt specific, well-defined conformations, we clustered the water-mediated contacts based on pairwise RMSDs (for details see the Methods section and Supplementary Material). The rationale behind this analysis is that, if water molecules do adopt well-defined specific positions in an interface, one might be able to derive for each type of water-mediated contact a few preferred conformations (an analogy in protein structures would be the rotameric states of side chains). Such information might be useful in the modelling of water-mediated contacts. The clustering statistics are reported in Supplementary Table 7. Using a 1.5 Å clustering cut-off almost 90% (118 out of 133) of the side chain contacts that could be clustered (133 out of 210) fall into one or two clusters (note that contacts for which less than two water-mediated instances were found could not be clustered at all). Figure 2 shows examples of clusters found for the most populated water-mediated contacts in the resolution-filtered Keskin dataset; in addition, the main backbone–backbone contact (O–H_2O–O) and the best-clustering backbone–side chain contact (Ser side chain–N) are shown.

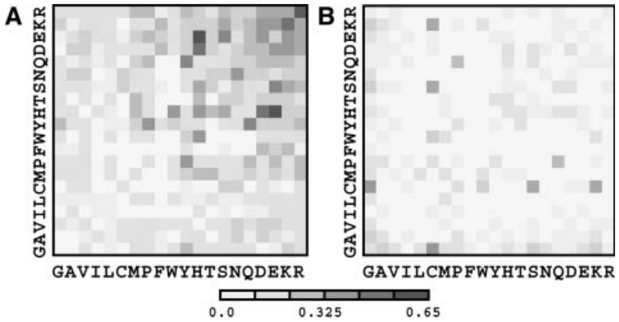

Fig. 1. Fraction of water-mediated contacts for each amino acid pairwise combination. Amino acid–amino acid contacts are colour-coded according to the fraction of water-mediated contacts for (**A**) side chain and (**B**) main chain contacts. The corresponding numbers for the matrix elements are provided as Supplementary Material (Supplementary Table 6).

3.2 Solvated docking

Our solvated docking approach is based on the concept of the first encounter complex in which the proteins are separated by a hydration layer. Before docking, we solvate the protein chains with one hydration layer as described in the Methods section. Then, the conventional HADDOCK rigid body docking protocol is followed; for this, each protein and its associated solvation shell is considered as one rigid body. This results in an encounter complex with a water-layer in between the two protein chains. All non-interfacial water molecules are removed from this complex and the remaining waters, together with the protein chains, are treated as separate rigid bodies in a subsequent energy minimization stage (1000 EM steps were found to be sufficient for convergence). Water molecules are then removed in a biased Monte Carlo procedure: randomly chosen water molecules are probed for their closest amino acid residues on both chains; their probability to be kept is set equal-to the observed fraction of water-mediated contacts for this specific amino acid combination as derived from the resolution-filtered Keskin set (see above). This procedure is repeated until only 25% of the initial interfacial water molecules remain. Subsequently, water molecules with an unfavourable interaction energy (sum of van der Waals and electrostatic water–protein energies >0.0 kcal/mol) are removed.

Finally, the remaining waters and the protein chains are again subjected to a rigid body energy minimization (for an overview see Supplementary Figure 6). Note that we checked that the use of water-mediated propensities to bias water removal does lead to improvement compared to a simple random removal of waters.

The number of retained waters at the end of our protocol is usually lower than 25% because of the energy criterion, typically between 10 and 20%. This fraction is roughly in accordance with a recent study (Rodier *et al*., 2005) where it was found that, on average, 90% of the interface waters are removed upon assembly. In fact, we observe a substantial variation in the final number of water

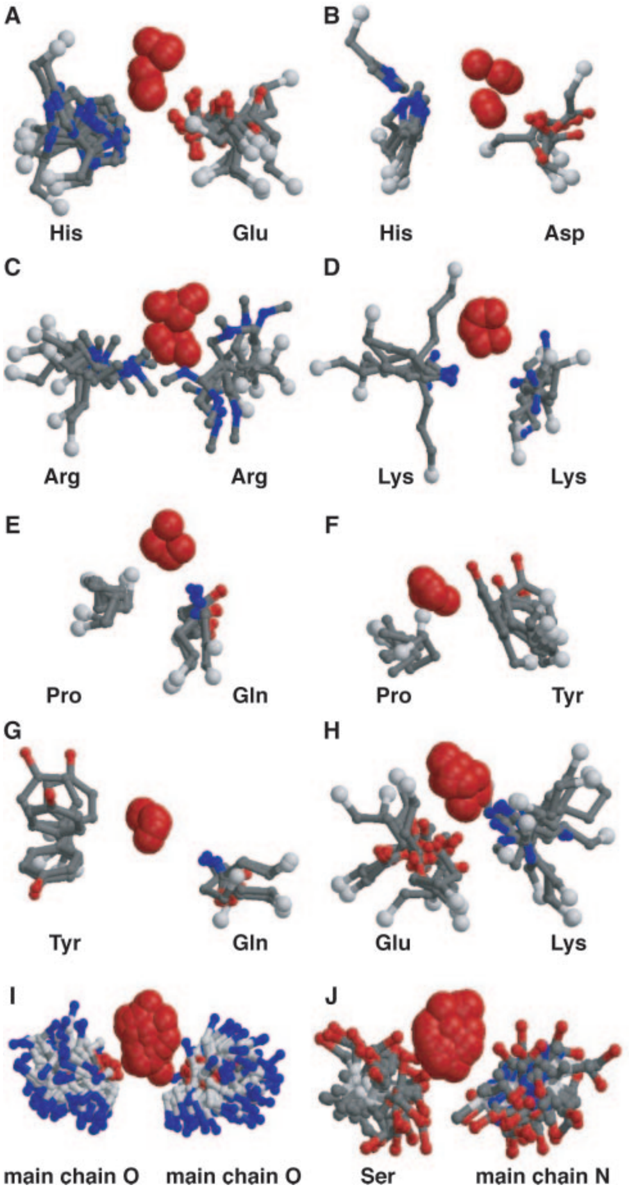

Fig. 2. View of the most populated water-mediated clusters as found for the resolution-filtered Keskin dataset. (**A**–**G**) Most often occurring side chain contacts; (**H**) best clustering side chain contact; (**I**) best clustering O–O main chain contacts; and (**J**) Ser side chain–main chain N contact. The clustering was performed based on positional RMSD (see the Methods section). Water oxygens are shown in red spheres. The amino acid side chain atoms are shown in ball-and-stick (red, oxygen; blue, nitrogen; grey, carbon), together with the Cα in white CPK. In the cases involving main chain contacts (I and J), only main chain atoms are shown.

molecules in the docked structures for the complexes that we used to test our protocol (see below, Table 4).

The solvated docking protocol as described above corresponds to the rigid body docking stage in HADDOCK. The resulting structures are then further refined using semi-flexible simulated annealing. Since water is introduced during rigid body docking we focus the discussion of our results on this stage, but we will also show some initial results for the semi-flexible refinement.

We tested our solvated docking approach on 10 complexes representing both 'wet' and 'dry' interfaces (Table 2). An accurate docking protocol accounting for the presence of water should not only be able to correctly position water molecules at the interface, thereby improving the docking results in the case of 'wet' interfaces, but also it should avoid retaining waters in 'dry' interfaces in order not to deteriorate the docking results. Assessed by the number of fully buried water molecules, the α-amylase–αAI and barnase–barstar complexes are representative of 'wet' interfaces, the PKC interacting protein complex represents a completely 'dry' interface and most of the other complexes are in-between. Only the E2A–HPr complex is an NMR structure for which no information on water positions is available.

The docking was performed using either the bound (B) structures from the complex or the unbound (U) structures; in the latter case rigid body docking was followed by flexible refinement. Experimental data (E) or interface residues (I) in the complex were used to define the AIRs, 50% of which were randomly discarded for each docking trial in the latter case (see the Methods section). Further details on these complexes and the information used to drive the docking can be found in the Methods section.

For each complex, two runs were performed: one reference run without water and one following our new solvated docking approach (see the Methods section). This was done for bound docking (using the bound structures of the components of the complex) and, if unbound structures were available, repeated for unbound docking.

The bound docking results are presented in Supplementary Table 8. Table 3 gives an overview of the unbound docking results, assessed by interface-RMSD (i-RMSD) to the target structure. The i-RMSD is defined as the backbone RMSD from the reference structure of the complex for those residues making contacts across the interface within a 10 Å cut-off [i-RMSDs below 2 and 4 Å are considered as medium quality and acceptable predictions, respectively, according to the CAPRI criteria (Mendez *et al.*, 2005)]. As can be seen from Table 3, the inclusion of water in docking generally improves the scoring of the solutions. This is clear from the i-RMSD of the top ranking solution: for the solvated docking, this is in five cases a medium quality solution and in one case an acceptable solution, whereas for the unsolvated docking, this is in only two cases a medium quality solution and in one case an acceptable solution. In addition, the rank of the best-ranked medium quality solution is in most cases lower for the solvated docking. Finally, the lowest RMSD found in all top 200 ranked structures is on average lower for the solvated docking. Note that scoring in our solvated docking protocol includes the water–water and water–protein non-bonded energy contributions, which clearly improves the performance (data not shown).

After flexible refinement (Table 3) the same conclusions are valid, although the differences between solvated and unsolvated docking are smaller. For example, the unsolvated docking has four medium and one acceptable solutions and the solvated docking has five medium quality solutions. For the 'wet' interfaces, a large fraction of the waters in our docking solutions have positions very close to those in the crystal (Fig. 3 and Supplementary Figures 7–9). These correspond to both fully buried waters and waters present at the rim of the interface. Especially the results from the bound barnase–barstar docking are impressive, with ~80% of the water molecules within 2 Å of crystal water positions. The distributions of distances

Table 2. Protein–protein complexes used in solvated docking

	PDB-id[a]	Res (Å)[b]	$N_{\mathrm{w,bur}}$[c]	BSA (Å^2)[d]	Docking[e]
α-amylase–αAI (BompardGilles *et al.*, 1996)	1dhk	1.9	25	3020	B/B; I
Barnase–barstar (Buckle *et al.*, 1994)	1brs; 1a2p; 1a19	2.0	18	1556	B/B + U/U; I
Subtilisin–subtilisin inhibitor (Takeuchi *et al.*, 1991)	2sic; 1sup; 3ssi	1.8	8	1617	B/B + U/U; I
Colicin E7–Im7 (Ko *et al.*, 1999)	7cei; 1ayi, 1cei, 1unk; 1m08	2.3	8	1384	B/B + U/U; I
bovine trypsin–CMTI-1 squash inhibitor (Bode *et al.*, 1989)	1ppe; 1btp; 1lu0	2.0	6	1688	B/B + U/U; I
Cohesin–dockerin (Carvalho *et al.*, 2003)	1ohz; 1anu	2.2	5	1504	B/B + U/B; E
GRB2 C-ter SH3 domain–N-ter SH3 domain (Nishida *et al.*, 2001)	1gcq; 1gcp; 1gri	1.7	4	1208	B/B + U/U; I
porcine trypsin–soybean trypsin inhibitor (Song and Suh, 1998)	1avx; 1ba7; 1qqu	1.9	1	1585	B/B + U/U; I
PKC interacting protein (Lima *et al.*, 1997)	1kpf	1.5	0	3700	B/B; I
E2A–HPr (Wang *et al.*, 2000)	1ggr; 1f3g; 1hdn	—[b]	—	1374	B/B+U/U; E

[a]PDB-id of the complex followed by the PDB-id of the unbound structures if available.

[b]Resolution; note that E2A–HPr (1ggr) is an NMR structure.

[c]Number of fully buried interfacial water molecules.

[d]Buried surface area as calculated using NACCESS (Hubbard and Thornton, 1993).

[e]The docking was performed using either the bound (B) structures from the complex or the unbound (U) structures; in the latter case rigid body docking was followed by flexible refinement. Experimental data (E) or interface residues (I) in the complex were used to define the ambiguous interaction restraints, 50% of which were randomly discarded for each docking trial in the latter case (see Methods).

Table 3. Unbound solvated and unsolvated docking results[a]

	Rigid body				Refined			
	Top 200[b] <4 Å	Top RMSD[c]	Best Rank[d]	Best RMSD[e]	All 200[b] <4 Å	Top RMSD[c]	Best Rank[d]	Best RMSD[e]
1brs								
R	**5**	**8.8**	**78***	**2.8**	**5**	**9.1**	**119***	**2.7**
S	**26**	**8.8**	**4**	**1.5**	**25**	**9.0**	**12**	**1.4**
2sic								
R	168	**1.9**	**1**	**1.6**	168	1.7	1	**1.3**
S	72	**1.9**	**1**	**1.5**	72	7.7	2	**1.3**
7cei								
R	**196**	**11.0**	**2**	**1.1**	**196**	1.3	**1**	0.8
S	**199**	**1.6**	**1**	**1.0**	**199**	1.4	**1**	0.9
1ppe								
R	198	**5.1**	**3**	**1.4**	198	1.1	**1**	**1.0**
S	186	**1.5**	**1**	**1.4**	186	1.2	**1**	**0.8**
1ohz (U/B)								
R	**19**	**5.8**	25	**0.7**	**17**	**6.2**	**38**	**1.1**
S	**33**	**3.1**	38	**0.7**	**33**	**6.0**	**2**	**0.7**
1gcq								
R	70	7.0	**19**	**1.4**	71	**4.0**	**2**	**1.1**
S	64	8.3	**3**	**1.4**	63	**1.7**	**1**	**1.1**
1avx								
R	194	1.4	**1**	1.4	194	1.6	**1**	1.0
S	171	1.7	**1**	1.5	171	1.9	**1**	1.1
1ggr								
R	106	**2.6**	**80**	**1.5**	106	**10.0**	**2**	**1.0**
S	96	**1.5**	**2**	**1.3**	95	**1.4**	**1**	**1.0**

[a]Results from reference (R) unsolvated and solvated (S) protein–protein docking for the various test cases (Table 2). Boldface indicates cases where solvated docking performs equal to or better than unsolvated docking.

[b]The number of structures below the indicated i-RMSD values is reported (<4 Å, acceptable quality). The i-RMSD is calculated over the backbone atoms of all residues making contacts across the interface within a 10 Å cut-off.

[c]i-RMSD of top ranking solution.

[d]Rank of best-ranked structure below 2 Å i-RMSD; when there are no structures below 2 Å i-RMSD this is the rank of the best-ranked structure below 4 Å i-RMSD (indicated with asterisk).

[e]i-RMSD value of best structure (closest to target).

between predicted and native waters in Figure 3 compare favourably with the results from Jiang *et al*. (2005); in that study, no docking was performed, but water positions at the interface were predicted from the crystal structures of a set of complexes. We also found that the quality of the water predictions does not change much after the semi-flexible refinement (Supplementary Figure 9). Note however that those are only preliminary results and the flexible refinement protocol needs further optimization.

We analysed the recovery of totally buried crystal water molecules over all acceptable (i-RMSD <4 Å) solutions out of the top 200 ranked models (Table 4 and Supplementary Table 9). On average, each docking solution contains between 6 and 12 water molecules (both buried and rim). Buried water molecules are generally more consistently recovered (i.e. found in a larger fraction of the solutions) than those at the rim of the interface (Fig. 4 and Supplementary Figures 9–11). On average, 94% of the buried crystal waters are recovered and each one is observed in 17% of the acceptable solutions. We find that those crystal waters that are not recovered are making most of their contacts with only one of the two components of the complex.

We also analysed the fraction of native water-mediated contacts recovered after flexible refinement: this is on average 30% for all acceptable structures, 46% for the highest-ranked acceptable structure and even 66% in the most favourable case. These are quite high fractions considering that on average, per structure, only 32% of the crystal waters are recovered within 4 Å. Those numbers are on average 25% smaller for rigid body docking solutions. As was already observed previously (van Dijk *et al*., 2005b), flexible refinement significantly improves the fraction of native contacts across the interface. In CAPRI, high/medium/acceptable-quality solutions require at least 50/30/10% fraction native contacts.

Crystal waters are recovered not only in 'wet' interfaces (e.g. α-amylase–αAI and barnase–barstar) but also, for example, in the case of 1gcq, where all four fully buried interface waters are found in several of the docking solutions [this complex shows the highest average fraction of structures in which crystal waters are observed (34%)]. For the 'dry' PKC interacting protein, the water molecules in the resulting docked structures are placed mostly at the rim of

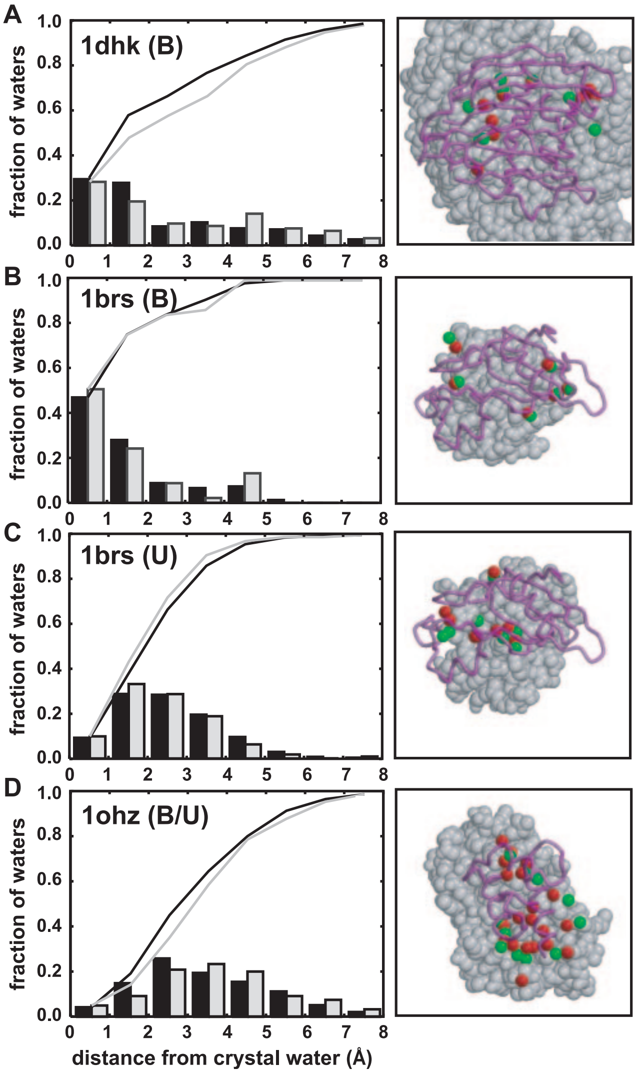

Fig. 3. Accuracy of predicted water molecules in solvated docking. (**A**) α-Amylase–αAI (bound docking); (**B**) barnase–barstar (bound docking); (**C**) barnase–barstar (unbound docking); and (**D**) cohesin–dockerin (bound/unbound docking). Left panel: Histograms (bars) and cumulative fractions (lines) of closest distances between modelled and crystal waters are shown for all acceptable structures (black) and for the top 10 acceptable structures (light grey) out of the top 200 ranked structures. Right panel: View of the best-scoring acceptable solvated docking solution, together with its predicted waters (red) and the corresponding ones in the crystal (green).

Table 4. Recovery of water molecules in solvated docking[a]

	<#waters>[b]	Rigid body Recovery[c]	# f_{recover} (%)[d]	Refined recovery[c]	# f_{recover} (%)[d]
1brs	10.4 (2.6)	14/18	12 (11)	14/18	14 (12)
2sic	11.7 (4.0)	7/8	29 (19)	7/8	29 (17)
7cei	8.2 (3.2)	8/8	5 (1)	8/8	3 (2)
1ppe	11.0 (3.7)	6/6	12 (10)	6/6	14 (9)
1ohz (U/B)	5.6 (4.6)	5/5	32 (11)	5/5	13 (8)
1gcq	10.9 (3.0)	4/4	34 (13)	4/4	28 (13)
1avx	8.9 (3.3)	1/1	10	1/1	11

[a]Solvated docking results for the acceptable solutions out of the top 200 models for the various test cases (Table 3). (B) and (U) indicate bound and unbound docking, respectively.

[b]Average number (standard deviation) of water molecules per structure.

[c]Number of fully buried crystal waters recovered (i.e. within 2.0 Å of a modelled water)/total number of buried crystal waters (Table 2).

[d]Average fraction (standard deviation) of acceptable structures in which a fully buried water is recovered.

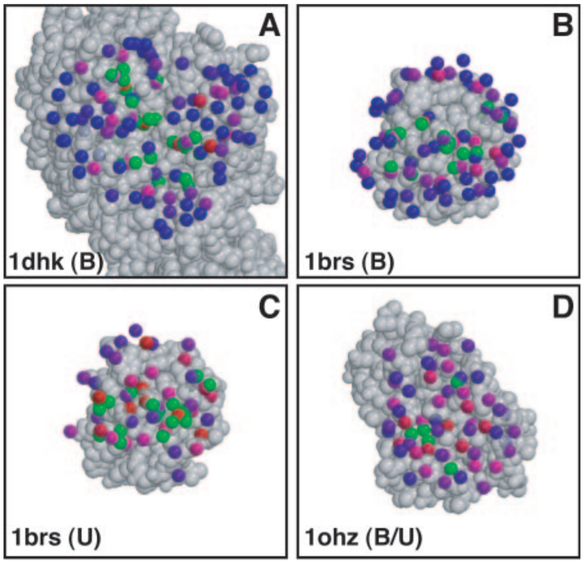

Fig. 4. Recovery of interfacial water molecules in solvated docking. (**A**) α-Amylase–αAI (bound docking); (**B**) barnase–barstar (bound docking); (**C**) barnase–barstar (unbound docking); and (**D**) cohesin–dockerin (bound/unbound docking). For each complex, the largest component is shown with its associated crystal waters (transparent green) together with cluster representatives from all predicted water in the acceptable solutions. The latter are colour-coded according to the fraction of acceptable structures in which they are observed, from blue 0% to red 40% (maximal observed fraction). Waters from all acceptable solutions were clustered based on pairwise distances using a 2.5 Å cut-off.

the interface. The same applies to E2A–HPr. For the latter, however, we cannot compare their positions to experimental ones since the reference complex was solved by NMR. Although decreasing somewhat the number of acceptable solutions for that particular complex, explicit inclusion of water led to an improvement in the ranking and in the number of medium quality solutions, both before and after flexible refinement. Taken all together, these results demonstrate the general applicability of our method.

Explicit inclusion of water molecules in our solvated docking protocol results in a factor 3 to 4 increase in computational time requirements for the rigid body docking stage. The most time-consuming part of HADDOCK is, however, the semi-flexible refinement stage, in which the presence of some additional water molecules does not make much difference. Explicit inclusion of water in docking thus only results in about a factor 2 increase in the overall run time, which is reasonable considering the improvements in both success rate and accuracy, and the fact that as a result water positions are predicted.

4 CONCLUSIONS AND PERSPECTIVE

For the first time, water has been introduced explicitly in protein–protein docking. We followed for this purpose a strategy mimicking the concept of the solvated initial encounter complex. By performing the docking from solvated protein chains in combination with a Monte Carlo water removal procedure based on water contact propensities, we successfully recovered interfacial crystal water molecules and improved our docking results both in bound and unbound docking cases. Further improvements could be achieved by making use of the geometrical information obtained from the cluster analysis of water-mediated contacts.

The very promising results obtained here and the rather reasonable additional computational burden make us confident that solvated docking is a viable approach to model biomolecular complexes. We actually started applying solvated docking in the last two rounds of CAPRI (targets 25 and 26; see http://capri.ebi.ac.uk) but will have to wait for the release of the targets in order to assess its performance. Solvated docking should also benefit the field of protein–DNA modelling since it is well known that protein–DNA complexes have rather wet interfaces. We therefore intend to extend our approach to the modelling of such complexes, which, as we demonstrated recently, can be modelled successfully using HADDOCK (van Dijk *et al*., 2006).

ACKNOWLEDGEMENTS

This work was supported by a ‘Jonge Chemici’ grant (grant no. 700.50.512) from The Netherlands Organization for Scientific Research (N.W.O.) to A.B. and by the European Community, FP6 STREP project ‘ExtendNMR’ (contract no. LSHG-CT-2005-018988).

Conflict of Interest: none declared.

REFERENCES

Babor,M. *et al*. (2002) Conserved positions for ribose recognition: importance of water bridging interactions among ATP, ADP and FAD–protein complexes. *J. Mol. Biol*., **323**, 523–532.

Bode,W. *et al*. (1989) The refined 2.0 a X-ray crystal-structure of the complex formed between bovine beta-trypsin and Cmti-I, a trypsin-inhibitor from Squash seeds (Cucurbita-Maxima)—topological similarity of the Squash seed inhibitors with the carboxypeptidase a inhibitor from potatoes. *FEBS Lett*., **242**, 285–292.

BompardGilles,C. *et al*. (1996) Substrate mimicry in the active center of a mammalian alpha-amylase: structural analysis of an enzyme–inhibitor complex. *Structure*, **4**, 1441–1452.

Boobbyer,D.N.A. *et al*. (1989) New hydrogen-bond potentials for use in determining energetically favorable binding-sites on molecules of known structure. *J. Med. Chem*., **32**, 1083–1094.

Brunger,A.T. *et al*. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr D*, **54**, 905–921.

Buckle,A.M. *et al*. (1994) Protein–protein recognition—crystal structural-analysis of a Barnase Barstar complex at 2.0-Angstrom resolution. *Biochemistry*, **33**, 8878–8889.

Carugo,O. (1999) Correlation between occupancy and B factor of water molecules in protein crystal structures. *Protein Eng*., **12**, 1021–1024.

Carugo,O. and Bordo,D. (1999) How many water molecules can be detected by protein crystallography? *Acta Crystallogr. D*, **55**, 479–483.

Carvalho,A.L. *et al*. (2003) Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proc. Natl Acad. Sci. USA*, **100**, 13809–13814.

Chandler,D. (2005) Interfaces and the driving force of hydrophobic assembly. *Nature*, **437**, 640–647.

Daura,X. *et al*. (1999) Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed*., **38**, 236–240.

Dominguez,C. *et al*. (2003) HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc*., **125**, 1731–1737.

Fernandez-Recio,J. *et al*. (2004) Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol*., **335**, 843–865.

Halperin,I. *et al*. (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, **47**, 409–43.

Houborg,K. *et al*. (2003) Impact of the physical and chemical environment on the molecular structure of *Coprinus cinereus* peroxidase. *Acta Crystallogr. D*, **59**, 989–996.

Hubbard,S.J. and Thornton,J.M. (1993) *NACCESS*. Department of Biochemistry and Molecular Biology, University College, London.

Hubbard,S.J. *et al*. (1994) Intramolecular cavities in globular-proteins. *Protein Eng*., **7**, 613–626.

Jiang,L. *et al*. (2005) A ‘solvated rotainer’ approach to modeling water-mediated hydrogen bonds at protein–protein interfaces. *Proteins*, **58**, 893–904.

Jorgensen,W.L. and Tirado-rives,J. (1988) The OPLS Potential functions for proteins. Energy minimizations for crystals of cyclin peptides and crambin. *J. Am. Chem. Soc*., **110**, 1657–1666.

Jorgensen,W.L. *et al*. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*., **79**, 926–935.

Kalodimos,C.G. *et al*. (2004) Structure and flexibility adaptation in nonspecific and specific protein–DNA complexes. *Science*, **305**, 386–389.

Keskin,O. *et al*. (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci*., **13**, 1043–1055.

Ko,T.P. *et al*. (1999) The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. *Structure*, **7**, 91–102.

Lima,C.D. *et al*. (1997) Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science*, **278**, 286–290.

Lindahl,E. *et al*. (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model*, **7**, 306–317.

Linge,J.P. *et al*. (2003a) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics*, **19**, 315–316.

Linge,J.P. *et al*. (2003b) Refinement of protein structures in explicit solvent. *Proteins*, **50**, 496–506.

Lo Conte,L. *et al*. (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol*., **285**, 2177–2198.

Loris,R. *et al*. (1999) Conserved water molecules in a large family of microbial ribonucleases. *Proteins*, **36**, 117–134.

Mendez,R. *et al*. (2005) Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins*, **60**, 150–169.

Mintseris,J. *et al*. (2005) Protein–protein docking benchmark 2.0: an update. *Proteins*, **60**, 214–216.

Moitessier,N. *et al*. (2006) Docking of Aminoglycosides to hydrated and flexible RNA. *J. Med. Chem*., **49**, 1023–1033.

Mustata,G. and Briggs,J.M. (2004) Cluster analysis of water molecules in alanine racemase and their putative structural role. *Protein Eng*., **17**, 223–234.

Nishida,M. *et al*. (2001) Novel recognition mode between Vav and Grb2 SH3 domains. *EMBO J*., **20**, 2995–3007.

Osterberg,F. *et al*. (2002) Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins*, **46**, 34–40.

Rarey,M. *et al*. (1999) The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins*, **34**, 17–28.

Raschke,T.M. (2006) Water structure and interactions with protein surfaces. *Curr. Opin. Struct. Biol*., **16**, 152–159.

Rashin,A.A. *et al*. (1986) Internal cavities and buried waters in globular proteins. *Biochemistry*, **25**, 3619–3625.

Rejto,P.A. and Verkhivker,G.M. (1997) Mean field analysis of FKBP12 complexes with FK506 and rapamycin: implications for a role of crystallographic water molecules in molecular recognition and specificity. *Proteins*, **28**, 313–324.

Robert,C.H. and Ho,P.S. (1995) Significance of bound water to local chain conformations in protein crystals. *Proc. Natl Acad. Sci. USA*, **92**, 7600–7604.

Rodier,F. *et al*. (2005) Hydration of protein–protein interfaces. *Proteins*, **60**, 36–45.

Schymkowitz,J.W.H. *et al*. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl Acad. Sci. USA*, **102**, 10147–10152.

Song,H.K. and Suh,S.W. (1998) Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from Erythrina caffra and tissue-type plasminogen activator. *J. Mol. Biol*., **275**, 347–363.

Sreenivasan,U. and Axelsen,P.H. (1992) Buried water in homologous serine proteases. *Biochemistry*, **31**, 12785–12791.

Takeuchi,Y. *et al*. (1991) Refined crystal-structure of the complex of subtilisin Bpn′ and Streptomyces Subtilisin inhibitor at 1.8 A-resolution.. *J. Mol. Biol*., **221**, 309–325.

Tame,J.R.H. *et al*. (1996) The role of water in sequence-independent ligand binding by an oligopeptide transporter protein. *Nat. Struct. Biol*., **3**, 998–1001.

van Dijk,A.D.J. *et al*. (2005a) Data-driven docking for the study of biomolecular complexes. *FEBS J*., **272**, 293–312.

van Dijk,A.D.J. *et al*. (2005b) Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins*, **60**, 232–238.

van Dijk *et al*. (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res*., **34**, 3317–3325.

Verdonk,M.L. *et al*. (2005) Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem*., **48**, 6504–6515.

Wade,R.C. and Goodford,P.J. (1993) Further development of hydrogen-bond functions for use in determining energetically favorable binding-sites on molecules of known structure.2. Ligand probe groups with the ability to form more than 2 hydrogen-bonds. *J. Med. Chem*., **36**, 148–156.

Wade,R.C. *et al*. (1993) Further development of hydrogen-bond functions for use in determining energetically favorable binding-sites on molecules of known structure.1. Ligand probe groups with the ability to form 2 hydrogen-bonds. *J. Med. Chem*., **36**, 140–147.

Wang,G. *et al*. (2000) Solution structure of the phosphoryl transfer complex between the signal transducing proteins HPr and IIA(glucose) of the *Escherichia coli* phosphoenolpyruvate:sugar phosphotransferase system. *EMBO J*., **19**, 5635–5649.

Yang,J.M. and Chen,C.C. (2004) GEMDOCK: a generic evolutionary method for molecular docking. *Proteins*, **55**, 288–304.

Zhang,X.J. and Matthews,B.W. (1994) Conservation of solvent-binding sites in 10 crystal forms of T4-Lysozyme. *Protein Sci*., **3**, 1031–1039.