This is the post-print version of the following article:

P.L. Kastritis, A.D.J. van Dijk and A.M.J.J. Bonvin
"Explicit Treatment of Water Molecules in Data-Driven Protein-Protein Docking: The Solvated HADDOCKing Approach"
Methods in Molecular Biology **819**, Part 5, 355-374 (2012)

Explicit treatment of water molecules in data-driven protein-protein docking: The solvated HADDOCKing approach

Panagiotis L. Kastritis[a], Aalt D J van Dijk[b] and Alexandre M.J.J. Bonvin[a*]

[a] Bijvoet Center for Biomolecular Research, Department of Chemistry, Science Faculty, Utrecht University, 3584CH, Utrecht, The Netherlands

[b] Applied Bioinformatics, Bioscience, Plant Research International, Wageningen UR, Droevendaalsesteeg 1, Wageningen, The Netherlands.

* To whom correspondence should be addressed. Prof. Dr. Alexandre M. J. J. Bonvin, NMR Spectroscopy Group, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands, E-mail: a.m.j.j.bonvin@uu.nl. Telephone: +310302533859. Fax: +31030253762.

Summary

Water molecules are active components in, literally, every biochemical event, forming hydrogen bonds, filling cavities and mediating interactions with other (bio)molecules. Therefore, solvent drastically affects the kinetics and thermodynamics of numerous cellular events, including protein-protein interactions. While docking techniques are becoming successful in predicting the three-dimensional structure of protein-protein complexes, they are still limited in accounting explicitly for water in the binding process. HADDOCK is one of the few programs so far that can explicitly deal with water molecules during docking. Its solvated docking protocol starts from hydrated molecules and a fraction of the interfacial water is subsequently removed from the docked models in a biased Monte Carlo procedure. The Monte Carlo-based removal is based on interfacial amino acid - water contact propensities derived from a dataset of high-resolution crystal structures of protein-protein complexes. In this chapter, this solvated docking protocol is described and associated methodological aspects are illustrated through an application example. It is shown that, although docking results do not always improve when the solvated docking protocol is applied, scoring is improved and the positions of buried water molecules in an interface are correctly predicted. Therefore, by identifying important water molecules, solvated docking can aid the development of novel inhibitors of protein-protein complexes that might be better accommodated at an interface.

Key words: protein complexes, HADDOCK, protein-protein docking, explicit model, solvation shell, Monte Carlo, structure prediction, solvated docking

1. Introduction

Water-protein interactions constitute a major determinant of the kinetics and thermodynamics underlying protein interactions (*1*). Over the last decades, advances in X-ray Crystallography (*2, 3*), Neutron Diffraction (*2*), Femtosecond Fluorescence (*4*), NMR spectroscopy (*5*) and Molecular Dynamics simulations (*6*) have opened the route to the establishment of methodologies for studying binding, structure and dynamics of water. These methods have revealed that water molecules are active components in, literally, every biochemical pathway, forming hydrogen bonds with the backbone or side chains of the polypeptidic chains, filling cavities and mediating interactions with other (bio)molecules.

Water molecules also play a key role in the hydrophobic effect in protein-protein binding. They can guide a fully solvated protein to recognize another fully solvated protein by a gradual expulsion of water layers. The water molecules that are finally trapped in an interface form hydrogen bonds that contribute to the enthalpy of binding while water molecules "released" from more apolar interfaces regain freedom in the bulk, resulting in an increase in entropy (*7*). In addition to the hydrophobic effect, water is a critical contributor to the specificity of protein-protein interactions: The wet nature of some protein-protein interfaces suggests that water is not randomly trapped in the interface, but is part of the recognition code, as it mediates interactions that would be less favorable in its absence (*8*). For example, water is a critical contributor to the cognate and non-cognate binding of

colicins and immunity proteins (*9, 10*), and completely different networks of water-mediated interactions are observed in the complexes of Barstar with Barnase (*11*) or RNAse S1 (*12*), respectively, resulting into dramatic differences in the binding affinities of those two complexes (*11, 12*).

Analysis of existing structures of protein-protein complexes has revealed an equal number of direct and water-mediated hydrogen bonds between the partner chains (*13*). Considering that (a) each water molecule in an interface can contribute ~1.5 kcal/mol to the total energy of the complex (*8*) and (b) their residence time is much longer (10-1000 ns) than that of other water molecules in the first hydration shell (~500 ps) (*5, 14*), buried waters should be considered as an integral part of the structure of a protein complex.

Computational modeling of the three-dimensional (3D) structure of biomolecular complexes, formed by two or more interacting biological macromolecules is referred to as macromolecular docking. When only proteins are considered, the term protein-protein docking is used. Docking typically consists of two different steps: the search through interaction space and the scoring of the resulting models. In the search step, a set of possible configurations for the 3D complex of interest are generated, typically starting from the free form structures of the partners that are being docked. The generated set should reliably include at least one nearly correct configuration (also termed “near native”). In the second step (scoring), the “near-native”, correct solutions have to be identified from the generated set of possible configurations of the complex.

Current docking methods have shown a substantial improvement throughout the years in predicting correctly the 3D structures of macromolecular complexes (*15, 16*). However, the role of water in both steps is, in most cases, ignored, contrary to the underlying physics of protein-protein association. During the search step, most of the docking algorithms consistently ignore the presence of water molecules, and, therefore, docking is performed in vacuum; even implicit representations of water are often ignored. Most of the algorithms include a desolvation term in the scoring function, which significantly improves the ranking of correct docking configurations (*17, 18*). Implicit treatment of the water comes with a price: approximations are introduced, and, compared to explicit models, the description of the energetics is coarser (*1, 19, 20*). In the standard HADDOCK protocol (*21, 22*), explicit waters are used in the final stage to refine models generated in vacuum. During this refinement, however, water molecules cannot diffuse into the interface to form specific contacts, but rather remain at the rim of the interface.

In this chapter, the solvated docking protocol implemented in our data-driven docking approach HADDOCK (*21, 22*) is discussed in details, demonstrating that water can be explicitly introduced in protein-protein docking. In the Theory section, the basic idea of the protocol is described, along with our docking program HADDOCK. The Method section explains how to actually perform a solvated docking calculation using the HADDOCK web server (*23*), and how to analyze and interpret the results. In the associated application section, we illustrate how docking results for Barnase, an extracellular ribonuclease, and Barstar, its intracellular inhibitor are improved when the standard solvated docking protocol (*24*) is applied: Water

molecules are recovered in all docking stages of HADDOCK and results from the explicit solvent refinement can be used to derive statistics about structural waters buried in the interface. In the Concluding Remarks and Perspectives section we discuss advantages and limitations of solvated docking along with potential applications.

2. Theory

The solvated docking protocol is a strategy that mimics the concept of the solvated encounter complex formed in the initial phase of protein-protein recognition. We perform the docking, starting from protein chains that are solvated in explicit shells of water molecules. Once the proteins have formed a 3D encounter complex, removal of water molecules trapped in the interface is achieved via a biased Monte Carlo (MC) approach. The latter is based on water-bridged amino acid – amino acid contact propensities derived from an analysis of high-resolution crystal structures of protein-protein complexes.

2.1. Residue-water contact propensities

The probabilities of finding water-mediated contacts in the interface are used to discard or keep waters in the initial stage of docking. These probabilities were derived from the non-redundant set of protein – protein complexes from Keskin *et al* (*25*).

For this analysis, interface residues were defined as residues having at least one heavy-atom contact with a residue from the partner chain within a 10Å distance cutoff. Water-mediated contacts were defined between pairs of interface residues, provided a water molecule is making at least one heavy-atom contact within 5Å with both residues. Propensities for residue pairs interacting with water molecules are shown in **Figure 1** (see **Note 1**). Probabilities for non-standard residue types or small molecules that appear in the interface are, in principle, unknown. However, an average interacting probability is assigned to them, using the average probability of the known elements of the matrix.

2.2. HADDOCK: High Ambiguity Data-driven DOCKing

HADDOCK is a molecular docking method driven by experimental knowledge in the form of information about the interface region between the molecular components and/or their relative orientations. In HADDOCK, experimental data are entered as active and passive residues. Identified interface residues are described as active residues, and their solvent accessible neighboring residues correspond to the passive ones. Active and passive residues are used to define a network of Ambiguous Interaction Restraints (AIRs) between the molecules to be docked. An AIR is defined as an ambiguous intermolecular distance (d_{iAB}^{eff}) 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_{resB} in total) of protein B (n_{kB}) (and inversely for protein A). The introduction of passive residues ensures that residues located in the interface but not detected (or predicted) can satisfy the

AIRs. The effective distance, corresponding to each restraint is calculated using the following equation:

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

where $\frac{1}{d_{m_{iA}n_{kB}}^{6}}$ denotes a potential that resembles the Lennard-Jones attractive term. The function has the property that d_{iAB}^{eff} will always be smaller than the shortest distance $d_{m_{iA}n_{kB}}$ entering the sum. The AIRs effectively enforce that the defined interfaces come together without imposing any restraint on their relative orientation.

2.3. Solvated docking

In our solvated docking protocol, the molecules to be docked are initially solvated in a shell of TIP3P water (*26*). Waters closer than 4.0Å or further away than 8.0Å from the protein surface are first removed. This results in a water layer surrounding each protein. Subsequently, a short Molecular Dynamics simulation is performed to optimize the water positions (see **Note 2**). After that, an additional removal of water molecules is performed, where only water molecules within 5.5 Å distance from the surface of the protein are kept. At this point, docking starts by rigid body minimization, during which each protein, with its corresponding solvation shell, is treated as one rigid entity. The resulting complex has two partly overlapping solvation shells (see **Figure 2**, after B step). All non-interfacial water molecules are

removed from the complex and the remaining waters, together with the protein chains are treated as separate molecules in a subsequent rigid body energy minimization stage.

Waters are then removed in a biased MC approach: water molecules are randomly picked and 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 water-mediated contact propensities (**Figure 1**). (see **Note 3**) The Monte Carlo process of interface water removal consists of the following steps:

1. A random water molecule is selected.
2. The distances between each water molecule and all neighboring atoms that belong to the first chain are calculated and the minimum distance for each water molecule is stored. The same is applied for the second chain (**Figure 3**, A).
3. The shortest distance interaction pair with its bridging water is assigned a probability to be kept that derives from the corresponding frequencies stored in a database file. The database file includes the pairing probabilities from the high-resolution structures originating from the Keskin dataset (*25*) (see **Note 4**).

This process (steps 1-3) is repeated, until a user-defined percentage (typically 25%) of the initial interfacial water molecules remain (**Figure 3,** A-C) (see **Note 5**)

4. Energetically unfavorable water molecules are removed that do not satisfy the criterion $E_{Elec}^{wat} + E_{vdW}^{wat} \leq 0$ (**Figure 3**, D) (see **Note 6**). The remaining

waters and the protein chains are again subjected to a final rigid body energy minimization, with each molecule treated as a separate rigid body.

The solvated docking protocol as described above corresponds to the rigid body docking stage in HADDOCK (see **Note 7**).

3. Method

Solvated docking with HADDOCK can also be performed using its web server implementation (*23*) (http://haddock.chem.uu.nl/services/HADDOCK) (**Figure 4**). In order to use the full functionalities of the web server and have full control over the solvated docking protocol, *guru interface* registration is required (see **Note 8**).

To fully understand the protocol that is described in this section it is highly recommended to read the articles describing the HADDOCK server and its usage (*23*) and the original work on solvated docking (*24*). Note that only significant parameters related to solvated docking will be discussed here. For unsolvated protein-protein docking consult other published material from our group (*21, 27, 28*). Do not alter other parameters unless otherwise stated.

3.1. Submitting a solvated docking calculation for the protein-protein complex of interest

1. Provided that you are registered as a guru user, go to http://haddock.chem.uu.nl/services/HADDOCK/haddockserver-guru.html

2. Unfold the menu of the '*First molecule*' and upload the PDB file of the first protein. This file should be in correct PDB format else the server will give an error. Note that the molecule can be directly downloaded from the PDB; just input the PDB ID of the molecule. This is, however, not recommended since it is better to inspect first and clean the files from unwanted/unnecessary molecules.
3. Define the chain that should be used for docking. If the protein consists of several chains that should all be used in docking, select the option 'all'. (see **Note 9**)
4. Define active and passive residues for the molecule. There are several experimental methods that can be used in order to define residues that are involved in protein-protein binding (*28-30*). For example, mutagenesis experiments, as well as Chemical Shift perturbation data from NMR experiments can be used as an input for the active residues. If no experimental information is available, docking can also be performed using bioinformatics interface predictions (for a review see (*29*)). To define active and passive residues, residue numbers should be inserted, corresponding to the number of the residues that are observed or predicted to be at the interface (see **Note 10**). For the passive residues, it is suggested to check the option to automatically define the passive residues related to the user-specified active residues (see **Note 11**).
5. Repeat steps 2-4 for the second protein molecule.
6. Turn on solvated docking under the Sampling Parameters box.

7. Unfold the Solvated Docking Parameters box. If the original protocol is to be followed, change the number of generated solvation shells from 1 to 5 (see **Note 12**). For more information on how to use alternative protocols of solvated docking, see **Note 13**.
8. Fill in your name and your password and submit.

3.2. Retrieving a docking run and analyzing the results

Once the HADDOCK run has finished, the results are accessible via a web link to the Results page, which has been automatically e-mailed to you. After a successful docking run, the clustered docking predictions will be displayed. Although the clusters are numbered according to cluster size, i.e. cluster 1 corresponds to the largest cluster and cluster 2 to the second largest, they are sorted by their HADDOCK score [see **Note 14**]. Only the top 10 clusters are displayed and the cluster with the lowest (best) HADDOCK score is on top of the web page, being the most plausible solution according to HADDOCK. For every cluster, the various components of the HADDOCK score are displayed. The top four structures of every cluster can be downloaded or viewed directly in a web browser using a JMol applet (http://jmol.sourceforge.net).

The entire docking run, containing all structures from all docking stages, is available for download in the form of a zipped tar archive. If the HADDOCK software has been installed on a local machine, the HADDOCK analysis and clustering steps can then be repeated with user-defined parameters. For this, download and save the archive in a local directory. Extract then the archive of the docking run; a new folder is created

with the same name as the specified run name. In this folder, final predictions from the solvated docking procedure can be found in *structures/it1/water*. PDB files including water molecules that derive from solvated docking share the same format: complex_X_h2o.pdb, where [ ] [see **Note 15**]. These underwent semi-flexible simulated annealing in torsion angle space, and final refinement in explicit solvent, according to the standard HADDOCK protocol. In order to visualize the models, a molecular graphics program is required (e.g. PyMol (http://www.pymol.org/)). [see **Note 16**].

3.3. Application example: Barnase and Barstar

Barnase is an extracellular nuclease that can interact very strongly with its cognate intracellular inhibitor Barstar, a protein with very high affinity and specificity for Barnase (*11*). Next to the well-established electrostatic steering of this interaction that guides the association of these proteins, water molecules play a critical role in the affinity and specificity of the interaction (*11*). The crystal structure of the protein-protein complex has been determined at 2.0Å resolution (*11*), revealing a very wet interface. Eighteen water molecules are found in a relatively small interface (1556Å^2), corresponding to the presence on average of one water molecule per 86Å^2 of the interface. Half of these waters correspond to bridging water molecules, forming in total 12 hydrogen bonds with residues from both chains.

Because using bound structures as starting point for docking does not correspond to the biological situation where unbound molecules bind to each other, the crystal structures of the unbound Barnase (PDB ID: 1A2P) (*31*) and Barstar (PDB ID: 1A19)

(*32*) will be used as input to predict the protein-protein complex. The positions of the water molecules in the interface will be predicted by solvated docking. The true interface definition is used [see **Note 17**] to focus on the role of buried interface water in the prediction of the protein-protein complex. To simulate a more realistic case, 50% of the restraints are randomly removed during docking. We follow the standard solvated docking protocol, described above. For comparison, a second docking run is performed, toggling off the solvated docking procedure but using otherwise the same settings. Results from both docking runs are evaluated according to the standard CAPRI criteria [see **Note 18**]. Generally, a high quality structure (***) means that the predicted complex closely resembles the true binding mode of the protein partners, a medium quality structure (**) corresponds to a reasonably good prediction, whereas an acceptable structure (*) indicates a near-native solution with correct interface, but with possibly some shift or rotation of the partner molecules. All other predictions that are not assigned a star are considered incorrect.

3.3.1. Results from unsolvated docking

When the proteins are docked using the standard HADDOCK protocol (unsolvated docking), a single cluster is generated. However, although nearly 400 docking solutions are of acceptable or better quality in the rigid body docking stage, only 46 are high quality predictions (***), ranking outside the top 200 structures in HADDOCK score (see **Table I**); medium-quality (**) predictions are also generated,

but still are not selected for the subsequent refinement stage, since they rank low. However, 83 acceptable predictions (*) rank very high within the top 200.

After semi-flexible refinement, 73 acceptable structures remain, the first one of rank 2. The final explicit refinement improves the structures substantially, leading to 108 acceptable predictions, corresponding to 54% of good predictions. Scoring is also good with the top 6 ranking structures all of acceptable quality. On average, acceptable structures rank much better than incorrect ones (see **Table I**).

3.3.2. Results from solvated docking

Three clusters of solutions were generated from the solvated docking run (see **Table II**). Although the first cluster is similar to the single cluster generated by the unsolvated docking protocol, two additional clusters are present. When results from solvated docking are retrieved and analyzed, high quality structures are now ranking at the top (see **Figure 5**). Even though the total number of acceptable or better structures generated in it0 is smaller compared to unsolvated docking (see **Table I**), scoring is greatly improved, leading to the selection of both high- (***) and medium- (**) quality predictions for semi-flexible refinement, whereas in unsolvated docking only acceptable- (*) quality structures were selected. After the semi-flexible refinement stage, 6 high quality structures are ranking at the top that can easily be selected from the pool of decoys. After final water refinement, one can observe a general improvement in the quality of the models, reaching 59% of acceptable or better solutions. The 3rd ranking (in terms of HADDOCK score) protein-protein complex that is generated, is a high quality (***) solution (see

Figure 5, A), resembling with high accuracy the bound conformation of Barnase-Barstar (PDB entry 1BRS)(*11*). It is included in the Top ranking cluster, which, on average, has a much better score than the other clusters generated in the solvated docking, or the single cluster from unsolvated docking (see **Table II**). Incorrect solutions after the final water refinement only appear at rank 9 or lower.

Such results clearly show that solvated docking can be used for protein-protein complexes in order to improve scoring. However, high quality data for the interface should be available in order to retrieve high quality results from the solvated docking and analyze conserved water positions with high confidence (see below and **Table III**).

3.3.3. Water positions in the generated solutions.

The crystal structure of Barnase and Barstar has in total 12 water-mediated hydrogen bonds, involving 9 bridging water molecules (*11*). All these water molecules and their interactions with the corresponding residues are well recovered in it0 (all of them are observed in the pool of acceptable solutions, although not consistently). However, after semi-flexible refinement some of those move to more energetically favored positions, e.g. forming contacts with residues that are both highly hydrophilic and can form a salt bridge. After the water refinement, the top-ranking cluster (see **Table III**) has a very high recovery rate of the water-mediated contacts observed in the crystal structure of the bound complex (1BRS), reaching >58% of correctly placed structural waters. The top ranking structure of the cluster is shown in **Figure 5, A**). We recommend, however,

analyzing all the members of the cluster to get reliable statistics about the position of structural waters (see also **Note 16**).

4. Concluding Remarks and Perspectives

The present application example is highlighting the direct influence of the water in the structure prediction of protein-protein complexes. A significant improvement in the quality of the docking predictions is observed compared to the standard unsolvated HADDOCK protocol for this system. Scoring of the models, as highlighted in the Barnase – Barstar example, can be improved when waters are explicitly accounted for during docking. On the other hand, when solvated docking is benchmarked on other systems (*24*) (tested initially in the original solvated docking development (*24*)), comparison with unsolvated docking results indicates that, for some of the complexes, scoring is improved and for others not. Since the original publication, HADDOCK has undergone over the years small but significant improvements that are reflected in a strong performance in the CAPRI competition (*21, 33*). Therefore, it should not be surprising that docking predictions with solvated docking are not always better, compared to the standard HADDOCK protocol.

Solvated docking can also be applied for structure prediction of multi-body protein complexes (*34*). The functionality has been already implemented in the multi-body docking interface of the HADDOCK webserver (*34*) (http://haddock.chem.uu.nl/services/HADDOCK/haddockserver-multi.html).

Experienced HADDOCK users can perform solvated docking with up to six biomolecules. However, the performance of solvated docking has not yet been benchmarked for complexes that are composed of more than two proteins, and therefore, it is suggested to use it with caution.

We are currently extending the knowledge-based probability method to account also for protein-DNA systems (Marc van Dijk, Utrecht University, *personal communication*). Solvated docking should be particularly important for these systems considering the wet interfaces of protein-DNA complexes. The presented solvated docking protocol can also easily be extended to protein-ligand docking: Although the pairing probabilities of a ligand are unknown, they are currently set to the average default value (see **Theory section 2.1**). New pairing probabilities involving pharmacophore groups could be derived from protein-ligand crystal structures deposited in the PDB. They could have a direct application in structure-based drug design for ligand optimization.

As a final remark, solvated docking can best be used in cases for which there is sufficient information about the interface. In such cases, the interface can be identified with confidence and our solvated docking protocol can predict fairly accurately the water positions. This is valuable information that can drive the development of novel inhibitors of protein-protein interactions by accounting for the structural role of waters at protein-protein interfaces, thereby increasing their specificity.

5. Notes

1. It is evident, as well as expected, that hydrophilic residues are observed to interact much more frequently with waters when compared to hydrophobic residues. These propensities drive the removal of the water molecules during docking. If a water molecule lays in-between two hydrophobic residues in the interface it is less likely to be kept, compared to water being in-between two hydrophilic residues.
2. The MD protocol consists of four times 1000 integration steps at 600, 500, 400 and 300K, respectively.
3. For example, a water molecule that is bridging a histidine and a glutamate is more likely to be retained (P(H,E) = 0.73) compared to a water molecule that is bridging two hydrophobic residues (e.g. isoleucine and valine, (P(I,V) = 0.08).
4. These probabilities vary ($P(Q) \in [0,\ 0.73]$). Therefore, the highest probability of a water molecule to be kept corresponds to the water molecule bridging E and H residues (see also **Figure 1**).
5. The fraction of interfacial water to be kept after the Monte Carlo removal process is an important parameter for the solvated docking protocol. Although it is set to 25% by default, water molecules that are kept in the interface could make unfavorable contacts and correspondly have a high energy. The cutoff percentage of 25, corresponds to the average percentage of the interface that is solvated from an analysis of protein-protein complexes (*8*)

6. Water molecules with unfavorable interaction energies (sum of van der Waals and electrostatic water–protein energies >0.0 kcal/mol) are, finally, removed. The number of retained waters at the end of the protocol is usually lower than 25% due to this energy criterion. In some cases, this criterion allows all interfacial water to be removed, which could be needed in the case of highly hydrophobic interfaces.
7. The resulting models, including the remaining water molecules, are then further refined using semi-flexible simulated annealing in torsion angle space, and final refinement of the derived complexes in explicit solvent, according to the standard HADDOCK protocol (*21, 22*).
8. There are three main web interfaces for HADDOCK, each corresponding to the experience level of the user: The *easy interface*, requiring only starting structures and lists of active and passive residues that will be used to drive the docking*, the expert interface*, allowing the more advanced user to upload custom restraints to drive the docking process, and *the guru interface*, providing full control over all aspects/parameters of the HADDOCK program.
9. If the option '*use all chains*' is selected, there should be no overlap in the residue numbering between the various chains of the molecule.
10. Residues that are considered active should be on the surface of the protein. We advise against setting all residues on the surface of the protein as active: next to increasing unnecessarily the computational time, they will result in large restraint violations, corresponding to very high energies of the resulting complexes.

11. This option assigns as passive, those residues that are both on the surface (relative surface accessibility of either main chain or side chain > 15%, as calculated with NACCESS (http://www.bioinf.manchester.ac.uk/naccess/)) and within a radius of 6.5Å of any active residue.

12. Generally, it is recommended to leave the solvated docking settings at their default values. In the original work, 5 different solvation shells were generated for each starting structure to assess the performance of the solvated docking protocol. If there are more than 1 starting structures for one of the proteins that are docked, you can leave this option to its default value.

13. Options '*initial cutoff for restrains solvating method*', '*cutoff for restraints solvating method*', '*Scale factor for restraints solvating method*', '*water-surface-cutoff*' should never be changed. These options correspond to another approach of solvated docking that has not yet been benchmarked. Briefly, water molecules are forced to be at close proximity to amino-acids that form the most water-mediated contacts (Arg, Asn, Asp, Gln, Glu, His, Lys, Pro, Ser, Thr and Tyr). This is done by defining ambiguous distance restraints between each water molecule and those amino acids on both sides of an interface. If '*fraction of water to keep in ntrial loop*' is changed, the fraction of water molecules that will be kept after the biased Monte Carlo removal procedure will be affected. By default it is 25% (therefore, boxes have the values 25 and 0.25). If more water molecules should be kept, these values must be set higher. Keep in mind that the protocol was tested to perform best

using default values. Finally, it is also possible to turn off water translation during rigid-body energy minimization if desired, disabling the option '*Use translation in loop miniwater*'. If the option '*Do some water analysis*' is selected, additional files will be generated with some water statistics. Note that when performing solvated docking via the web server interface additional PDB files with extension *_.h2o.pdb* will be written in the *structures/it1/water* directory. These contain both the complex and the water molecules.

14. The HADDOCK score (given in arbitrary units) cannot be used to predict binding affinities or compare different complexes (*35*). It should only be used to compare different solutions for a given complex. The reported scores are evaluated on the top 4 members of a given cluster.

15. The PDB files in the water directory do not contain the standard chainID (column 22) that distinguishes the chains in a complex. This information is instead stored in what is called the SegID (columns 73-76). Since most molecular viewers use the chainID to distinguish between chains it is convenient to transfer first the SegID information to the ChainID. Provided a local version of HADDOCK has been installed, this can be done with the following command: `$HADDOCKTOOLS/pdb_segid-to-chain input-pdb > output-pdb`

16. In order to have good statistics for the water positions, more than one model should be analyzed. For example, to derive which water molecules are found in the interface and are conserved throughout the docking run, a large majority of the structures present in the (top ranking) cluster should be

analyzed. We are currently developing analysis scripts that will be included in the *tools* directory of the downloadable docking archive in a future version of HADDOCK.

17. The true interface is defined as those residues directly involved in the interaction of the partners. Interface residues of barnase that served as input for HADDOCK were 35, 37, 38, 55-60, 62, 73, 82-84, 101-104 and 106, whereas interface residues for barstar were 27, 29, 30, 31, 33-36, 38-40, 42-47, 73 and 76. This information was converted into Ambiguous Interaction Restraints (AIRs) via the GenTBL page of the HADDOCK website (http://www.nmr.chem.uu.nl/sevices/GenTBL) and the generated file was uploaded directly in the distance restraints section of the server.

18. Docking decoys are evaluated using the ligand root mean square deviation (L-RMSD), interface RMSD (i-RMSD) and fraction of native contacts (f_{nat}).

    The classification is as follows:

    ***, high quality prediction: $f_{nat} \geq 0.5$ and (L-RMSD ≤ 1.0 or i-RMSD ≤ 1.0)

    **, medium quality prediction: $f_{nat} \geq 0.3$ and (L-RMSD ≤ 5.0 or i-RMSD ≤ 2.0)

    *, acceptable quality prediction: $f_{nat} \geq 0.1$ and (L-RMSD ≤ 10.0 or i-RMSD ≤ 4.0)

ACKNOWLEDGEMENTS

This work was supported by the Netherlands Organization for Scientific Research (VICI Grant 700.56.442 to **AMJJB** and VENI Grant 863.08.027 to **ADJvD**) and the European Community (FP6 integrated Project SPINE2-COMPLEX Contract 032220 and FP7 e-Infrastructure “e-NMR” I3 project, Grant 213010)

REFERENCES

1. Levy, Y., and Onuchic, J. N. (2006) Water mediation in protein folding and molecular recognition, *Annu Rev Biophys Biomol Struct* ***35***, 389-415.
2. Savage, H., and Wlodawer, A. (1986) Determination of water structure around biomolecules using X-ray and neutron diffraction methods, *Methods Enzymol* ***127***, 162-183.
3. Halle, B. (2004) Biomolecular cryocrystallography: structural changes during flash-cooling, *Proc Natl Acad Sci U S A* ***101***, 4793-4798.
4. Pal, S. K., and Zewail, A. H. (2004) Dynamics of water in biological recognition, *Chem Rev* ***104***, 2099-2123.
5. Otting, G., Liepinsh, E., and Wuthrich, K. (1991) Protein hydration in aqueous solution, *Science* ***254***, 974-980.
6. Park, S., and Saven, J. G. (2005) Statistical and molecular dynamics studies of buried waters in globular proteins, *Proteins* ***60***, 450-463.
7. Petrone, P. M., and Garcia, A. E. (2004) MHC-peptide binding is assisted by bound water molecules, *J Mol Biol* ***338***, 419-435.
8. Rodier, F., Bahadur, R. P., Chakrabarti, P., and Janin, J. (2005) Hydration of protein-protein interfaces, *Proteins* ***60***, 36-45.
9. Cascales, E., Buchanan, S. K., Duche, D., Kleanthous, C., Lloubes, R., Postle, K., Riley, M., Slatin, S., and Cavard, D. (2007) Colicin biology, *Microbiol Mol Biol Rev* ***71***, 158-229.
10. Meenan, N. A., Sharma, A., Fleishman, S. J., Macdonald, C. J., Morel, B., Boetzel, R., Moore, G. R., Baker, D., and Kleanthous, C. (2010) The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction, *Proc Natl Acad Sci U S A* ***107***, 10080-10085.
11. Buckle, A. M., Schreiber, G., and Fersht, A. R. (1994) Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution, *Biochemistry* ***33***, 8878-8889.
12. Sevcik, J., Urbanikova, L., Dauter, Z., and Wilson, K. S. (1998) Recognition of RNase Sa by the inhibitor barstar: structure of the complex at 1.7 A resolution, *Acta Crystallogr D Biol Crystallogr* ***54***, 954-963.
13. Bahadur, R. P., and Zacharias, M. (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions, *Cell Mol Life Sci* ***65***, 1059-1072.
14. Denisov, V. P., and Halle, B. (1995) Protein hydration dynamics in aqueous solution: a comparison of bovine pancreatic trypsin inhibitor and ubiquitin by oxygen-17 spin relaxation dispersion, *J Mol Biol* ***245***, 682-697.
15. Lensink, M. F., Mendez, R., and Wodak, S. J. (2007) Docking and scoring protein complexes: CAPRI 3rd Edition, *Proteins* ***69***, 704-718.
16. Lensink, M. F., and Wodak, S. J. (2010) Docking and scoring protein interactions: CAPRI 2009, *Proteins* ***78***, 3073-3084.
17. Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2004) Identification of protein-protein interaction sites from docking energy landscapes, *J Mol Biol* ***335***, 843-865.

18. Fernandez-Recio, J., Abagyan, R., and Totrov, M. (2005) Improving CAPRI predictions: optimized desolvation for rigid-body docking, *Proteins* ***60***, 308-313.
19. Zhou, R. (2003) Free energy landscape of protein folding in water: explicit vs. implicit solvent, *Proteins* ***53***, 148-161.
20. Snow, C. D., Sorin, E. J., Rhee, Y. M., and Pande, V. S. (2005) How well can simulation predict protein folding kinetics and thermodynamics?, *Annu Rev Biophys Biomol Struct* ***34***, 43-69.
21. de Vries, S. J., van Dijk, A. D., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A. M. (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets, *Proteins* ***69***, 726-733.
22. Dominguez, C., Boelens, R., and Bonvin, A. M. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information, *J Am Chem Soc* ***125***, 1731-1737.
23. de Vries, S. J., van Dijk, M., and Bonvin, A. M. (2010) The HADDOCK web server for data-driven biomolecular docking, *Nat Protoc* ***5***, 883-897.
24. van Dijk, A. D., and Bonvin, A. M. (2006) Solvated docking: introducing water into the modelling of biomolecular complexes, *Bioinformatics* ***22***, 2340-2347.
25. Keskin, O., Tsai, C. J., Wolfson, H., and Nussinov, R. (2004) A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications, *Protein Sci* ***13***, 1043-1055.
26. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water, *J Chem Phys* ***79***, 926-935.
27. Fuentes, G., van Dijk, A. D., and Bonvin, A. M. (2008) Nuclear magnetic resonance-based modeling and refinement of protein three-dimensional structures and their complexes, *Methods Mol Biol* ***443***, 229-255.
28. van Dijk, A. D., Boelens, R., and Bonvin, A. M. (2005) Data-driven docking for the study of biomolecular complexes, *FEBS J* ***272***, 293-312.
29. de Vries, S. J., and Bonvin, A. M. (2008) How proteins get in touch: interface prediction in the study of biomolecular complexes, *Curr Protein Pept Sci* ***9***, 394-406.
30. Melquiond, A. S. J., and Bonvin, A. M. J. J. (2010) Data-driven docking: Using external information to spark the biomolecular rendez-vous, in *Protein-protein complexes: Analysis, modeling and drug design* (Zacharias, M., Ed.), pp 182-208, Imperial College Press, London.
31. Martin, C., Richard, V., Salem, M., Hartley, R., and Mauguen, Y. (1999) Refinement and structural analysis of barnase at 1.5 A resolution, *Acta Crystallogr D Biol Crystallogr* ***55***, 386-398.
32. Ratnaparkhi, G. S., Ramachandran, S., Udgaonkar, J. B., and Varadarajan, R. (1998) Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable, *Biochemistry* ***37***, 6958-6966.
33. de Vries, S. J., Melquiond, A. S., Kastritis, P. L., Karaca, E., Bordogna, A., van Dijk, M., Rodrigues, J. P., and Bonvin, A. M. (2010) Strengths and weaknesses of

data-driven docking in critical assessment of prediction of interactions, *Proteins 78,* 3242-3249.

34. Karaca, E., Melquiond, A. S., de Vries, S. J., Kastritis, P. L., and Bonvin, A. M. (2010) Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server, *Mol Cell Proteomics* ***9***, 1784-1794.
35. Kastritis, P. L., and Bonvin, A. M. (2010) Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark, *J Proteome Res* ***9***, 2216-2225.

FIGURE CAPTIONS

Figure 1. Calculated propensities for pairs of amino acids interacting with water. Amino acids are sorted according to the Kyte-Doolittle hydrophobicity scale.

Figure 2. Schematic of solvated docking steps. (A) Short MD run in a solvation shell to optimize the water positions. Water molecules far from the protein (>5.5Å distance) are subsequently removed. (B) Rigid-body docking of the proteins with the optimized water layers. (C) Removal of non-interfacial water and energy minimization (for more, see Theory section) (D) Biased Monte Carlo removal of interfacial waters and further removal of energetically unfavourable interface waters based on their corresponding energetics (waters are removed when $E_{vdW}^{wat} + E_{Elec}^{wat} > 0$) and final minimization.

Figure 3. The biased Monte Carlo procedure illustrated through an example consisting of 8 water molecules: (A) Interfacial water molecules are randomly selected and their corresponding minimum distance from residues in the interacting chains are identified. (B) according to the probabilities that were derived (see Figure 1), water molecules are either kept or discarded. (C) When only 25% of the water molecules are remaining, the MC procedure is stopped and, (D) an energetic criterion is applied to further remove unfavorable water molecules.

Figure 4. The 'guru' interface of the HADDOCK web server, providing full control over all HADDOCK parameters and supporting all experimental restraints that can drive the docking procedure. The solvated docking field is shown.

Figure 5. Best docking results (dark grey color) from unsolvated (left) and solvated (right) docking, using unbound barnase and barstar superimposed on top of the bound complex (light grey – PDB ID: 1BRS). Best docking results refer to the model from the top ranking cluster with the lowest i-RMSD from the crystal structure. On the bottom, a comparison of the interfacial waters found in the crystal structure (PDB entry 1BRS)(*11*) (dark grey) and recovered by solvated docking (light grey) are shown (the hydrogen atoms are not shown). Residues contacting these water molecules are represented as sticks.

TABLES

TABLE I. Docking results for the Barnase-barstar complex, in terms of quality of the generated structures[1].

Docking stage		Unsolvated docking				Solvated docking			
	Quality	***	**	*	**UN**	***	**	*	**UN**
it0	**Number**	46	196	156	602	44	172	197	587
	Best rank	649	201	2	1	168	18	5	1
it1	**Number**	0	0	73	127	25	7	53	115
	Best rank	n/a	n/a	2	1	1	12	8	6
water	**Number**	0	0	108	92	30	2	87	81
	Best rank	n/a	n/a	1	7	3	130	1	9

[1]Stars correspond to the standard CAPRI quality criteria, 'UN' denotes unacceptable docking predictions. It0, it1 and water correspond to the (a) rigid-body minimization, (b) semi-flexible simulated annealing in torsion angle space and (c) explicit solvent refinement stages of HADDOCK, respectively. Number rows correspond to the number of structures of different quality generated at each docking steps (total number of generated models is 1000, 200, 200 for it0, it1 and water, respectively.). Rank is the best ranking structure from each corresponding category (the lower this number, the better), generated at each stage.

TABLE II. Energetics and statistics of the generated Clusters in solvated and unsolvated docking. The best ranking cluster is highlighted in bold.

	Cluster Rank	**Size**	***i-RMSD [Å]***	f_{nat}	**HADDOCK Score [a.u.]**	**van der Waals energy [kcal/mol]**	**Electrostatic energy [kcal/mol]**	**Desolvation term [a.u.]**
Solvated docking	**1**	***34***	***2.0±0.7***	**0.65±0.15**	**-170±21**	**-54.8±8.7**	**-364.3±69.3**	**-3.9±9.2**
	2	*20*	*12.1±0.9*	0.08*±0.02*	-159±17	-63.2± 9.1	-218.7±42.0	-0.7± 7.1
	3	*141*	*10.9±0.3*	0.10*±0.02*	-152±22	-56.5 ±8.2	-245.9±54.2	-8.4±8.1
Unsolvated docking	**1**	**200**	**10.9*±0.3***	**0.10*±0.02***	**-109±15**	**-58.0±7.5**	**-238.2±47.4**	**-13.4±7.4**

TABLE III. Specific water molecule recovery for the best cluster of the solvated docking run[2] .

Water-mediated contacts observed in the crystal structure		**Water-mediated contacts observed in the best cluster**
Barnase	**Barstar**	
Lys62	Asp35	**+(6)**
Asn58	Asp35	**+(8)**
Arg59	Asp35	**+(4)**
Tyr103	Asp35	**-**
Ile55	Trp38	**-**
Glu73	Trp38	**-**
Lys27	Asp39	**+(2)**
Glu73	Asp39	**+(5)**
Arg83	Gly43	**+(2)**
Ser38	Val45	**-**
Ser38	Tyr47	**-**

[2]Contacts on the left are present in the crystal structure of the barnase-barstar complex (PDB entry 1BRS) as reported in the original manuscript (*24*). (+) and (-) represent the presence or absence of the water-mediated contact in the best cluster. The numbers in parentheses represent their frequencies (cluster size=34)

Figure 1

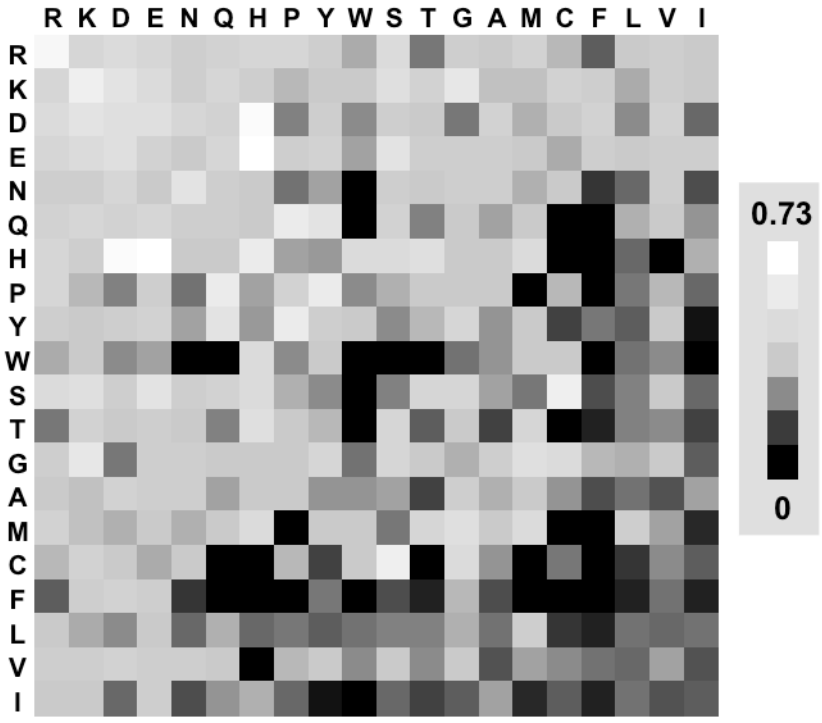


Figure 2

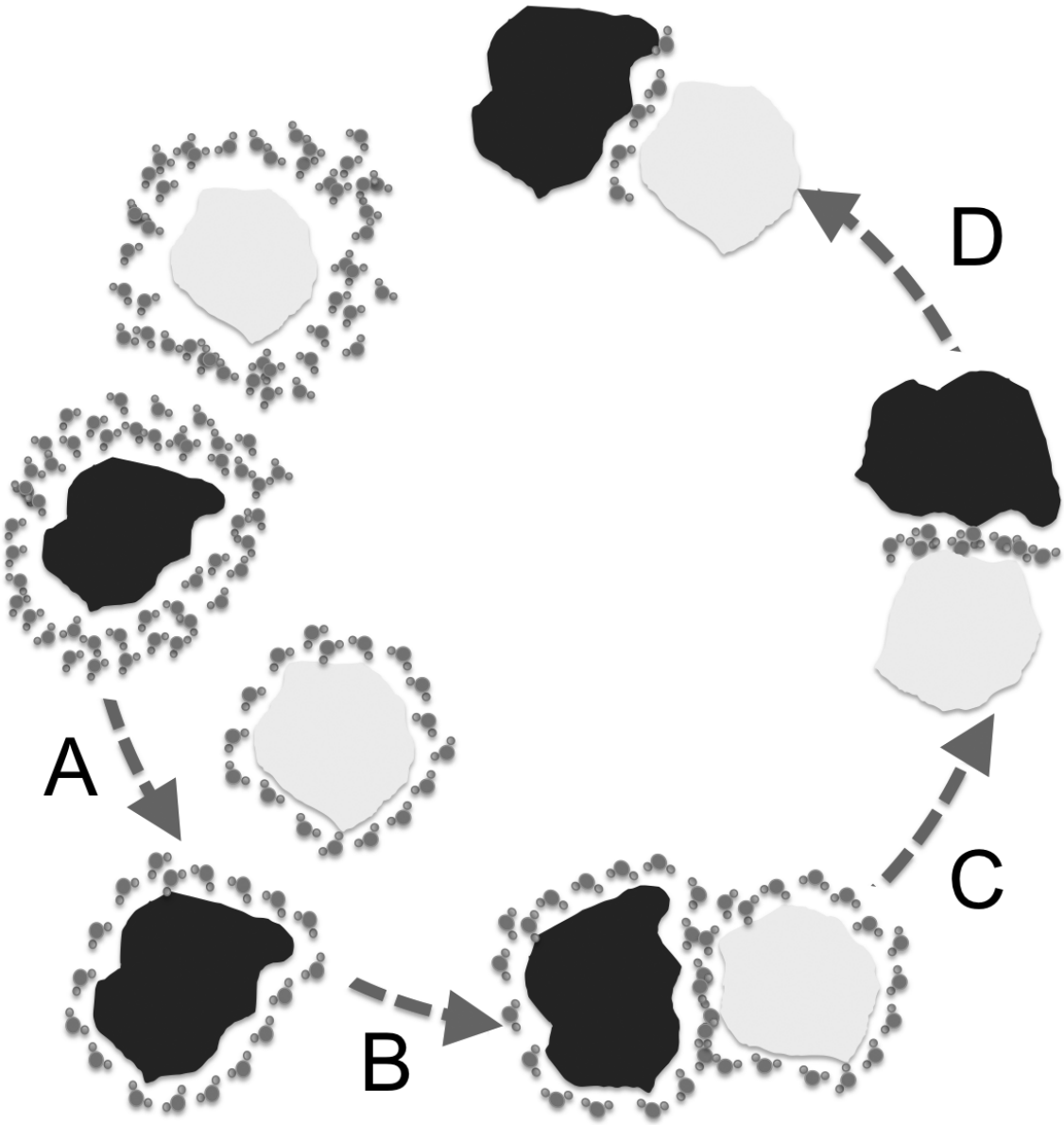


Figure 3

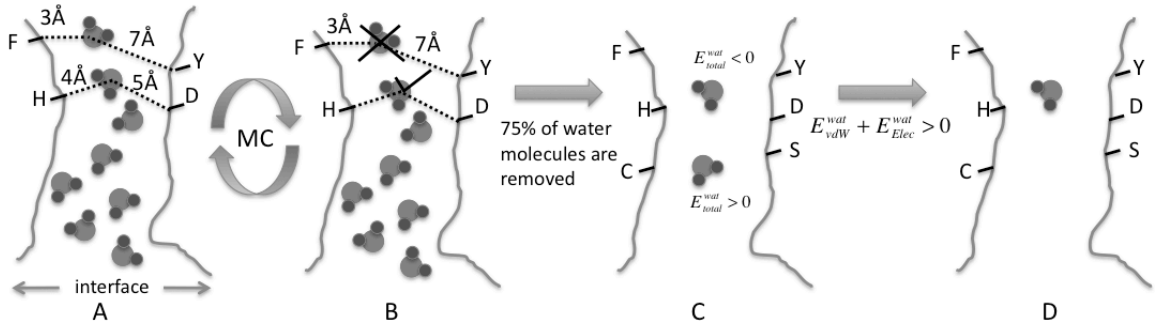

3Å
7Å
4Å
5Å
F
Y
H
D
MC
interface
A
B
75% of water molecules are removed
C
$E^{wat}_{total} < 0$
$E^{wat}_{total} > 0$
S
$E^{wat}_{vdW} + E^{wat}_{Elec} > 0$
D

Figure 4

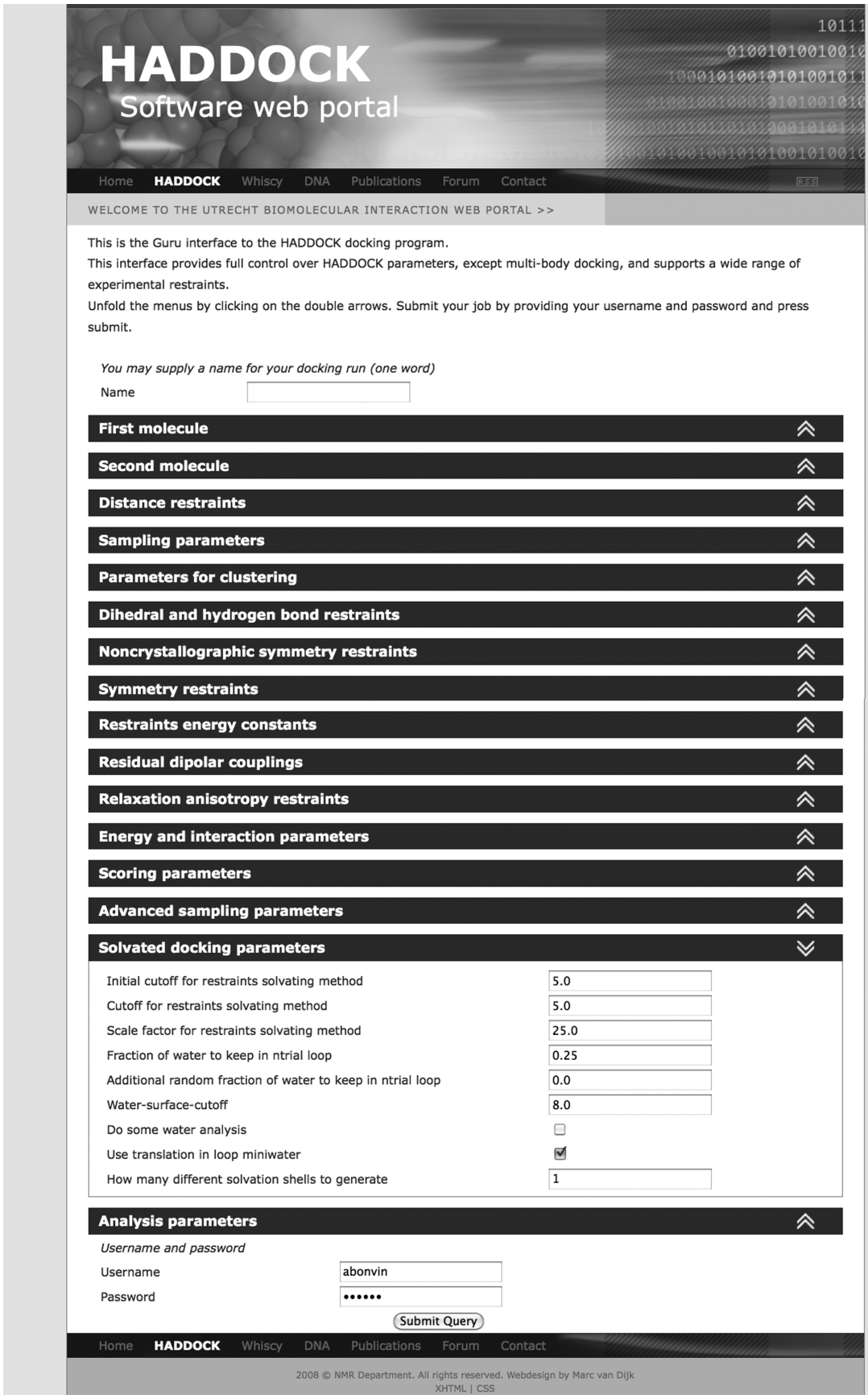
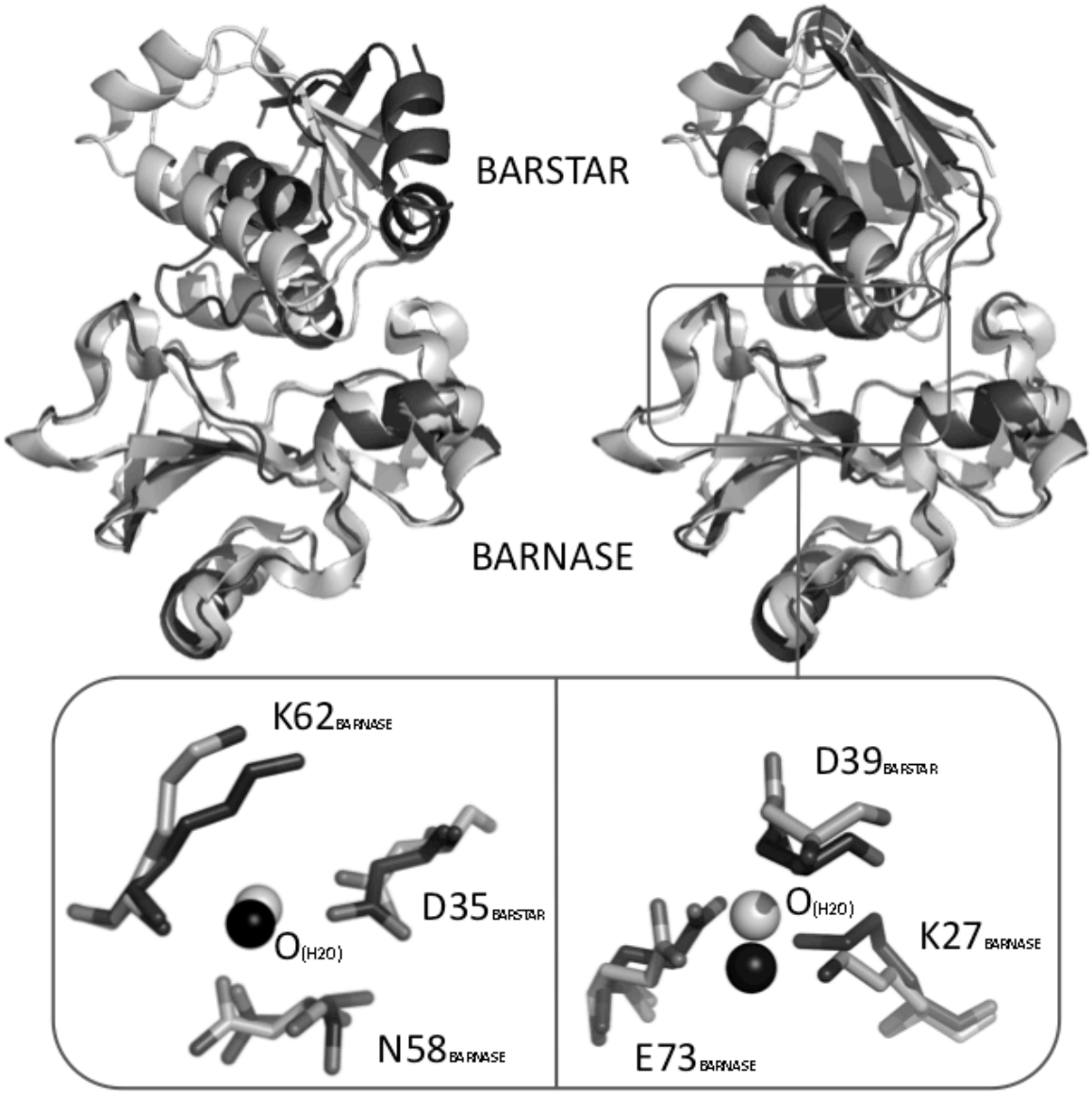

HADDOCK
Software web portal
Home HADDOCK Whiscy DNA Publications Forum Contact
WELCOME TO THE UTRECHT BIOMOLECULAR INTERACTION WEB PORTAL >>
This is the Guru interface to the HADDOCK docking program.
This interface provides full control over HADDOCK parameters, except multi-body docking, and supports a wide range of experimental restraints.
Unfold the menus by clicking on the double arrows. Submit your job by providing your username and password and press submit.
You may supply a name for your docking run (one word)
Name
First molecule
Second molecule
Distance restraints
Sampling parameters
Parameters for clustering
Dihedral and hydrogen bond restraints
Noncrystallographic symmetry restraints
Symmetry restraints
Restraints energy constants
Residual dipolar couplings
Relaxation anisotropy restraints
Energy and interaction parameters
Scoring parameters
Advanced sampling parameters
Solvated docking parameters
Initial cutoff for restraints solvating method 5.0
Cutoff for restraints solvating method 5.0
Scale factor for restraints solvating method 25.0
Fraction of water to keep in ntrial loop 0.25
Additional random fraction of water to keep in ntrial loop 0.0
Water-surface-cutoff 8.0
Do some water analysis
Use translation in loop miniwater
How many different solvation shells to generate 1
Analysis parameters
Username and password
Username abonvin
Password
Submit Query
Home HADDOCK Whiscy DNA Publications Forum Contact
2008 © NMR Department. All rights reserved. Webdesign by Marc van Dijk
XHTML | CSS

Figure 5


BARSTAR
BARNASE
K62 BARNASE
D35 BARSTAR
O(H2O)
N58 BARNASE
D39 BARSTAR
O(H2O)
K27 BARNASE
E73 BARNASE