Solvated protein-DNA docking using HADDOCK

Marc van Dijk[1], Koen Visscher[1], Panagiotis L. Kastritis[1] and Alexandre M.J.J. Bonvin[1*]

[1] Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands.

[*] To whom correspondence should be addressed. Email: a.m.j.j.bonvin@uu.nl, Phone: +31.(0)30.2533859, Fax: +31.(0)30.2537623

Keywords: Complexes, interface, water, protein, DNA

Abstract

Interfacial water molecules play an important role in many aspects of protein-DNA specificity and recognition. Yet they have been mostly neglected in the computational modeling of these complexes. We present here a solvated docking protocol that

allows explicit inclusion of water molecules in the docking of protein-DNA complexes and demonstrate its feasibility on a benchmark of 30 high-resolution protein-DNA complexes containing crystallographically-determined water molecules at their interfaces. Our protocol is capable of reproducing the solvation pattern at the interface and recovers hydrogen-bonded water-mediated contacts in many of the benchmark cases. Solvated docking leads to an overall improvement in the quality of the generated protein-DNA models for cases with limited conformational change of the partners upon complex formation. The applicability of this approach is demonstrated on real cases by docking a representative set of 6 complexes using unbound protein coordinates, model-built DNA and knowledge-based restraints. As HADDOCK supports the inclusion of a variety of NMR restraints, solvated docking is also applicable for NMR-based structure calculations of protein-DNA complexes.

Introduction

In the last decade, numerous genomics (Collins et al. 2003; Zhang et al. 2011), proteomics (Pandey and Mann 2000; Renuse et al. 2011) and interactomics (Collura and Boissy 2007; Cusick et al. 2005) efforts have enriched our understanding of the biomolecular world we live in. A substantial part of that knowledge has been contributed by the structural biology field, which is adding the structural dimension to these efforts and providing an atomistic view on biomolecules and their interactions. Proteins that interact with DNA play an important role in the context of interaction networks by regulating many cellular processes involving gene expression, DNA replication and repair.

The amount of structural information concerning protein-DNA complexes is increasing rapidly due to numerous efforts from the X-ray crystallography (Pakotiprapha and Jeruzalmi 2009) and NMR spectroscopy (Clore 2011; Varani et al. 2004) fields and to computational advances in analysis (Dunn and Kingston 2007), modeling (Aloy and Russell 2006; Baker and Sali 2001; van Dijk and Bonvin 2010) and simulations (Giudice and Lavery 2002; Pérez et al. 2012). In particular docking, a computational approach that models the unknown structure of a complex from its constituents, is a valuable tool to study complex formation in interaction networks (Melquiond et al. 2011). In the context of protein-DNA complexes, docking has been used for screening potential interaction partners (Roberts et al. 2004), studying specific interactions (Liu and Bradley 2012) and assisting at various stages of the experimental workflow (Chen et al. 2008).

As in many computational techniques, also in docking there is a tradeoff between available computational resources and the ability to answer scientific questions with sufficient details (Samish 2009). With the focus on adequately and quickly sampling

the relevant conformational space, docking is, in most cases, performed *in vacuo*, neglecting the physical, aqueous, environment where the biomolecules are functional. However, in the last years it has become apparent that water molecules play an active role in nearly all aspects of biomolecular recognition and interaction (Ahmad et al. 2011; Ball 2008; Janin 1999; Li and Lazaridis 2007). For protein-DNA interactions in particular, water molecules are involved in diverse tasks such as screening for favorable DNA interaction sites, stabilizing complex formation and facilitating specific interactions (Janin 1999; Jayaram and Jain 2004; Reddy et al. 2001; Schwabe 1997). For example, the specificity in the *trp* repressor-operator complex is governed nearly exclusively by water-mediated amino acid to base interactions (Otwinowski et al. 1988; Shakked et al. 1994). Despite increasing computational resources, it is surprising that water molecules are still mostly neglected in docking protocols. Only few applications have been reported, limited to the prediction of solvation patterns in known complexes or their constituents (Virtanen et al. 2010), the inclusion of interfacial water molecules in NMR structure calculations of a protein-DNA complex (Kalodimos et al. 2004) and the inclusion of water molecules in the docking of protein-ligand and nucleic acid-ligand complexes (Huang and Shoichet 2008; Moitessier et al. 2006). The first two applications model the water molecules after the complex has been formed, thus neglecting the possible effect water has on the complex formation process. For the purpose of docking this is irrelevant because the structure of the complex is not yet known. It is thus important that explicit water molecules are present during complex formation, an approach successfully applied to the docking of ligand molecules as mentioned above. We successfully applied a similar approach to the docking of protein-protein complexes in our previously reported solvated protein-protein docking protocol implemented in HADDOCK

(Kastritis et al. 2012a; Kastritis et al. 2012b; van Dijk and Bonvin 2006). In this protocol, the protein chains are solvated in a primary layer of explicit water and subsequently docked. In proceeding from the initial encounter complex to the final structure, the excess of interfacial water molecules are removed in a biased Monte Carlo procedure based on interfacial hydrophobicity or water-mediated contact propensities. The protocol leads to improvements in both quality and scoring with respect to *in vacuo* docking and was able to recover many of the water molecules observed in the reference crystal structures.

In this study we describe the adaptation of this method to the solvated docking of protein-DNA complexes demonstrating the successful inclusion of explicit water during the docking of these complexes. We extended the protocol by including protein-DNA specific water-mediated contact propensities derived from statistical studies performed by Marabotti *et al.* (2008) and Luscombe et al. (2001). The target fraction of interfacial water molecules after biased removal was doubled to 0.5 based on the observation that an average protein-DNA complex contains up to twice as many interfacial water molecules as a protein-protein complex (Jones et al. 1999). Finally we enabled multi-body solvated docking, since many protein-DNA complexes consist of more than two molecules.

This protocol is tested on a benchmark of 30 high-resolution protein-DNA complexes showing a successful prediction of the interface solvation pattern, recovery of many hydrogen bonded water-mediated contacts and improvement in the overall quality and scoring for many of the generated models. As a demonstration of the applicability of this protocol, we also present solvated docking results for 6 representative real cases, starting from unbound protein structures and DNA partners using experimentally-derived ambiguous interaction restraints (AIRs). These have been previously docked

in a non-solvated setting (van Dijk and Bonvin 2010).

Results

Our solvated protein-DNA docking protocol (described in details in the Methods section) was tested on a new benchmark of 30 high-resolution protein-DNA crystal structures. The benchmark was constructed in a similar manner as our previously published protein-DNA benchmark (van Dijk and Bonvin 2008) but considering only crystal structures with a resolution of 2 Å or better giving higher confidence in the resolved, crystallographically-determined interfacial water molecules and their positions (see Methods section). The resulting protein-DNA benchmark (**Table 1**) is composed of a diverse set of complexes with respect to their mode of interaction (according to the classification of Luscombe *et al*. (Luscombe et al. 2000)), to the amount of protein conformational changes upon complex formation and to their interface solvation patterns. Unbound DNA structures are not included but instead the DNA in the bound conformation is used for docking. The benchmark contains complexes with various degrees of interface water molecules, from "dry" (1cdw) to very "wet" interfaces (1a73). Within the region we classify as interface (see Methods section) interface water molecules are often differently distributed, being either fully buried in the interface or positioned along the rim often following the DNA sugar-phosphate backbone (Jayaram and Jain 2004; Li and Lazaridis 2007; Reddy et al. 2001; Spyrakis et al. 2007) (**Table 1**). Examples of such unequal distribution are observed in DNA minor groove binding proteins of the TBP family (TATA-box binding protein: 1cdw, 1qne, 1ytb) that interact with the minor groove of the DNA where the double helix is splayed open and curves away from the protein. These complexes have little water at the core of the interface as an excess of water in the

minor groove of the TATA box is thermodynamically expensive (Dunitz 1994). Instead, most of the water molecules align along the rim of the interface, stabilizing the complex and dampening the electrostatic repulsion of the negatively charged phosphate backbone (Nadassy et al. 1999). Restriction enzyme cases, on the other hand are classified as having "wet" interfaces. In those, the many interface water molecules are proposed to play an important role in enzyme specificity (Horton 1998; Schwabe 1997; Sidorova and Rau 1996).

Because of the similarity with our previously published benchmark (van Dijk and Bonvin 2008), we make this new solvated protein-DNA docking benchmark available as a merged, updated version (1.3) at http://haddock.science.uu.nl/dna/benchmark.html

The 30 structures in the benchmark were docked using the solvated and the regular non-solvated protocols of HADDOCK. This allows evaluating both the effect of explicit solvation on the quality of the docking models and the recovery of interfacial water molecules with respect to the reference complex. We first docked the protein(s) and the DNA in their bound state using true interface derived restraints to minimize the effect of conformational change upon complex formation and of the quality of the data used to construct the AIRs used to drive the docking. This allows us to focus exclusively on the recovery of interfacial water molecules and defines the best-case scenario. By subsequently docking the protein(s) in their unbound state to the DNA (the latter in its bound state), we evaluate the effect of explicit solvation on the quality of the generated models considering only conformational change on the protein side.

Effect of interfacial water molecules on the quality of the docking models

HADDOCK successfully reconstructed the interface of the complex in 80% of the bound-bound docking cases, leading to acceptable or better solutions according to CAPRI quality criteria (Lensink and Wodak 2010a) in the selected 20 solutions (see Methods section). The remaining cases (1bgb, 1eyu, 1rva, 2oaa, 2odi, 3v6t) involve protein structures that adopt a closed conformation around the DNA after complex formation. For such cases, we previously successfully used an approach where non-bonded interactions were scaled down to 1% in the initial rigid body docking stage, allowing interpenetration of the docking partners resulting in CAPRI medium to high quality solutions (van Dijk and Bonvin 2010). For solvated docking the same approach did not result in any improvement mainly due to the steric hindrance of the many water molecules at the interface that create a complex energy landscape resulting in low quality solutions with very dry interfaces. Therefore, we did not consider these cases in the further analysis of the docking results.

We evaluated the effect of interfacial water molecules on the interface RMSD (i-RMSD) and fraction of native contacts (fNAT) as model quality descriptors for the best 20 docking solutions based on the HADDOCK score after water refinement and clustering (**Table 2**). The i-RMSD and fNAT show that HADDOCK is capable of reconstructing the interface with high accuracy using both solvated and non-solvated docking in a bound-bound docking setting. Nearly all models score as high quality, sub-ångstrom 3-star models according to the CAPRI criteria. Although the improvements in iRMSD and fNAT in the sub-ångstrom range are limited, the solvated docking protocol significantly improves the quality of the docking models in 55% of the benchmark cases; 42% remains unchanged while only 2 cases significantly degraded in quality when applying solvated docking. The latter two docking models included intricate interfaces were multiple protein structural elements

interact with the DNA grooves and/or other protein interfaces. The many water molecules trapped at these interfaces likely prevent accurate solutions.

This trend is also observed in solvated bound-unbound docking where significant improvements are apparent for 42% of the benchmark cases while 17% remain unchanged. Conformational change between unbound and bound conformations of the protein(s) often prevent the formation of high quality models as is the case for bound-bound docking leading to 41% that do not benefit from solvated docking. Still, solvated docking improved the quality for a number of cases that undergo considerable conformational change upon complex formation. Overall, solvated docking will not make the difference in the quality of the models expressed in the number of CAPRI stars, but it will improve the accuracy of the predicted interfaces in many cases, a feature that is most prominent for cases that closely resemble the native interface.

Recovery of protein-DNA interface solvation patterns

A successful solvated docking protocol should be able to reproduce the differences in interface solvation patterns that exist among different types of protein-DNA complexes. **Figure 1** shows the correlation between the observed numbers of interfacial water molecules in the best 20 docking solutions with respect to the reference structures for bound-bound (**Fig. 1a**) and bound-unbound docking (**Fig. 1b**). The number of interfacial water molecules is reported as those fully buried in the interface and those located at the rim of the interface (see the Methods section on how these two regions are determined). The data are also given in supplementary **Table S1**.

For bound-bound docking there is a significant correlation between experimentally observed and predicted waters, both for the total number of interfacial water molecules and those fully buried, indicating that the solvation patterns that exists in these complexes are well reproduced. This is further corroborated by the observation that ~20% of all the recovered water molecules are positioned within 1.5Å of a water oxygen atom in the reference crystal complex after fitting on the interface (**Table 2**).

For bound-unbound docking the same significant correlation is found for the total number of interfacial water molecules, but the number of fully buried water molecules retained is smaller than for bound-bound docking. This may originate from the flexibility of the residues at the interface that could cause the fully buried water molecules to be expelled to the rim region during docking, leading to a different distribution of interfacial water molecules. The conformational changes at the interface and their possible effect on the distribution of interfacial water molecules make it difficult to objectively compare the positions of recovered water molecules with those in the reference structure. We therefore did not perform such analysis for the bound-unbound docking cases.

Figure 1c and **1d** illustrate the resemblance between top ranking bound-bound docking models and the reference structure for the TATA-box binding protein (1qne) and the Intron-encoded homing endonuclease I-PPOI (1a73). These are examples of "dry" and "wet" interfaces, respectively, as described in the discussion of the benchmark. The figures clearly illustrate the ability of the protocol to reproduce the differences in distribution of interfacial water molecules. Overall, these results show that our method is able to efficiently recover the overall solvation level of the interfaces and the differences in interface solvation patterns that exists between fully buried and interface rim water molecules. These are most accurately predicted for

cases that closely resemble the native interface (i.e. showing limited conformational changes upon binding). A tendency towards a less solvated interface core is observed when conformational change occurs at the interface.

Recovery of water-mediated hydrogen bonded contacts

Many of the interfacial water molecules in protein-DNA complexes establish hydrogen bonded contacts with the protein, the DNA or both, acting as mediators in the formation of intermolecular hydrogen bonds (Reddy et al. 2001). **Figure 2** shows that the total number of hydrogen-bonded contacts was well recovered in the best 20 docking solutions with respect to the reference crystal structure for both bound-bound (**Fig. 2a**) and bound-unbound docking (**Fig. 2b**). The role of these water molecules in protein-DNA recognition and complex formation is diverse as reviewed previously (Jayaram and Jain 2004). We classified the hydrogen bonded interfacial water molecules into those only contacting the DNA or those facilitating water-mediated contacts between amino acids and the DNA sugar-phosphate backbone or bases. These three classes are consistently populated in both bound-bound and bound-unbound docking for all benchmark cases, with respectively 50 ± 20%, 44 ± 19% and 5 ± 7% of the water molecules recovered. This is consistent with previous studies that have reported that around 80% of the hydrogen bonded interfacial water molecules are involved in DNA backbone hydration (Nadassy et al. 1999). Their putative role is to reduce electrostatic repulsions and thus act as an electrostatic buffer around the highly charged DNA backbone.

In terms of recovery of hydrogen-bonded water-mediated contacts (f^{w}_{nat}), most bound-bound benchmark cases score in the "fair" category with a considerable percentage in the "good" and "excellent" range as well, according to CAPRI quality

criteria defined for water recovery in target 47 (see http://www.ebi.ac.uk/msd-srv/capri/round23/) For bound-unbound docking this distribution shows a shift towards the "fair" and "bad" regions for the majority of the cases. However, considering that the total number of interfacial water molecules and the number of hydrogen bonded water molecules is reasonably well recovered and that the majority of the hydrogen-bonded contacts are involved in non-specific backbone hydration, the shift for bound-unbound docking is likely due to water molecules making different but functionally equivalent contacts. Although the frequency of specific amino acid to DNA base contacts is far less than the non-specific DNA backbone contacts, our solvated docking protocol is still able to recover specific contacts of biological interest. For example, specific water-mediated contacts between Asn50 – Ala30 and Lys49 – Gly29 in the Drosophila engrailed homeodomain (2hdd) for instance, are recovered in 50% of the best solutions (**Fig. 2d**).

Solvated protein-DNA docking using unbound partners and experimentally derived restraints

In most day-to-day HADDOCK applications the bound conformations of the docking partners are unknown and the restraints used to drive the docking are knowledge-based. To demonstrate the applicability of our solvated docking protocol under 'real-life docking' conditions we performed solvated docking on the same 6 representative test cases previously used to validate our two-stage protein-DNA docking protocol (van Dijk and Bonvin 2010). In that study, in order to allow for larger conformational changes in the DNA, a two-stage docking protocol was followed. In brief this protocol starts out with the unbound coordinates of the protein and an ideal canonical model of the DNA. The partners are docked using AIRs defined based on biochemical

and biophysical information from literature sources. The results of the first docking round are then analyzed with respect to trends in the conformational change in the DNA after complex formation. This information is subsequently used to generate new pre-bend and twisted DNA 3D structural models (van Dijk and Bonvin 2009) used in a second docking round. We applied our new solvated protein-DNA docking protocol to this second docking round using a pre-bent DNA library.

The iRMSD and fNAT quality descriptors (**Table 3**) shows that solvated docking significantly improves the quality in the case of complexes undergoing limited conformational changes upon binding (the *'Easy'* cases, ΔiRMSD ≤ 2Å), and this is also observed for the more challenging intermediate cases (ΔiRMSD 2-5Å) of the protein-DNA benchmark (van Dijk and Bonvin 2008). For the most challenging cases undergoing extensive (ΔiRMSD > 5Å) conformational changes upon complex formation, the *'Difficult'* category of the benchmark, solvated docking did not result in any significant improvement, but also did deteriorate the results significantly.

Discussion

We have demonstrated here the feasibility of introducing explicit water molecules in the modeling of protein-DNA systems using HADDOCK. We extended our previously reported protein-protein solvated docking approach that mimics the concept of the solvated initial encounter complex and expanded it to deal with protein-DNA systems.

The modified protocol successfully recovers specific solvation patterns and water-mediated contacts observed in many of the interfaces of the diverse set of complexes in our protein-DNA benchmark. The benefits of our approach on the overall quality

and information content of the generated docking models in comparison to non-solvated docking are most apparent for those cases in which the unbound docking partners adopt a conformation close to their bound state. This becomes evident from the various bound-unbound cases and the 'easy' and 'intermediate' unbound-unbound docking runs. Furthermore, DNA interacting proteins that create intricate interfaces in which multiple structural elements interact with the DNA major and/or minor grooves are less likely to benefit from solvated docking. The success rate of solvated docking is not influenced by differences in scoring and clustering of the solutions. For both solvated and non-solvated docking, the HADDOCK score and fraction of common contact clustering (Rodrigues et al. 2012) were able to select the best solutions among the selected 20 models. The benchmark cases for which solvated docking was less successful predominantly show a lower recovery of fully buried interfacial water molecules, often in the major or minor grooves, indicating that these water molecules are either removed from the interface or expelled to the rim region during docking and flexible refinement. The procedure used by HADDOCK to generate the initial solvation shell is not designed to recreate the typical hydration spine and ribbon observed in the DNA minor- and major grooves respectively and other approaches to generate the initial solvation shell could be considered (Giudice and Lavery 2002; Hummer et al. 1995; Makarov et al. 2002). Since HADDOCK can keep the water molecules already present in the initial structure, ordered water molecules from molecular dynamics simulations for instance could be used to provide a more realistic DNA hydration pattern to start the docking with. Altogether, the improvements in overall model quality due to solvated docking will not make the difference between having acceptable or higher quality solutions, it will however add another level of information to the model that can aid in understanding specific details of complex

formation and molecular recognition in a protein-DNA complex, with possible implications as well for drug design. The practical use of our solvated protein-DNA protocol thus becomes most evident for "refinement" docking, a docking setting in which little conformational change is needed to accurately assemble the interface. The statistical power of (water mediated) contact analysis performed on clustered results of these docking runs will also increase both the confidence and the information content of the models. We expect the performance of solvated docking to improve with the amount and quality of the information available to define the protein-DNA complex. As such it can thus also be applied in classical structure determination by NMR, where intermolecular protein-DNA NOE restraints can drive the docking. The solvated protein-DNA docking protocol will be integrated into the upcoming version 2.2 of the HADDOCK docking web portal (http://haddock.science.uu.nl).

Methods

Solvated protein-DNA docking benchmark

The solvated protein-DNA docking protocol was validated using a non-redundant benchmark composed of 30 structures deposited in the RCSB Protein Data Bank (PDB, as of April 2012) (Berman et al. 2007). The PDB was queried for all entries resolved by x-ray crystallography with a resolution ≤ 2.0 Å containing protein and double-stranded DNA but not RNA/DNA hybrids or Z-type DNA. Entries contain ligands, modified polymeric residues or mutations in core and/or interface regions as well as double-stranded DNA structures with a length less then one helical turn (~10 base-pairs) where removed. For structures with a sequence similarity ≥ 90%, the entry with the highest structural completeness and/or highest resolution was selected. For the resulting complexes, the PDB was queried for unbound protein entries resolved by

Nuclear Magnetic Resonance (NMR) spectroscopy or X-ray crystallography. All entries for which no unbound equivalent was found were removed. The final set of 30 structures of the complexes and equivalent unbound proteins was cross-referenced to the PDB_Redo (Joosten et al. 2012) and RECOORD (Nederveen et al. 2005) databases. Refined X-ray and NMR structures from these structure recalculation efforts were used if they showed improvements in terms of X-ray or NMR quality criteria and general structure validation assessments as reported by these databases.

Restraints used to drive the docking

True interface Ambiguous Interaction Restraints (AIRs). HADDOCK uses restraints to drive the docking and as such they are a determinant for the quality of the generated models by influencing the correct assembly of the complex and driving potential conformational changes in the flexible stages of the docking. Usually these restraints are knowledge-based, derived from various biochemical and/or biophysical sources (van Dijk et al. 2005). Gathering enough data of sufficient quality for all of the 30 complexes of the benchmark is not only difficult but would lead to a bias in docking performance which would hamper an objective assessment of the performance of the protocol. Therefore docking was performed instead with true interface-derived ambiguous interaction restraints. These were defined based on the interfaces of the reference complexes as defined by the residues involved in intermolecular atom–atom contacts ≤ 5.0 Å. Contacts that originated from amino-acid residues having a relative main-chain or side-chain solvent accessibility < 30% as measured by NACCESS (Hubbard and Thornton 1993) where discarded. All residues used in creating the restraints were defined as 'active'. During docking, 50% of the restraints where discarded at random for each docking trial. Although not needed for

true interface derived restraints, in the case of experimental information random removal allows correcting for false positives in noisy data sets. Effectively we used the same procedure to generate and use AIRs as in the case of experimental information with the difference that they are only defined between the residues that are known to be in close vicinity in the reference complex. Note that the Ambiguous Interaction Restraints in HADDOCK do not define the relative orientation of the molecules, but only force the defined interfaces to come together.

Docking protocol

The default protein–DNA docking protocol (van Dijk and Bonvin 2010; van Dijk et al. 2006) implemented in HADDOCK (de Vries et al. 2007; Dominguez et al. 2003) (High Ambiguity Driven DOCKing) version 2.2 using CNS (Brunger 2007) version 1.3 was used for all the docking runs. The solvated protein-protein docking extension (Kastritis et al. 2012a; van Dijk and Bonvin 2006) published before was used to develop the protein-DNA optimized equivalent as explained briefly in the following.

Topology generation. The docking partners were immersed in a box of TIP3P (Jorgensen et al. 1983) water molecules. All water molecules outside a cut-off range (< 4.0 Å to > 8.0 Å) from the protein or DNA were removed. A short molecular dynamics (MD) run was performed to optimize the water positions while keeping the proteins or DNA fixed (4000 MD steps consisting of four times 1000 steps at a temperature of 600, 500, 400 and 300 K, respectively). Finally, all water molecules at a distance higher than 5.0 Å from the protein or DNA were removed.

Rigid-body docking (it0). 2000 rigid-body docking solutions were generated. Every docking was performed 5 times and for each the symmetrical 180° rotated solution was also sampled; out of those 10 docking trials, the solution with the lowest

HADDOCK score was saved to disk. For solvated docking, the initial encounter complex has a water layer trapped between the partners. All non-interfacial water molecules are first removed from this complex. The remaining water molecules, together with the protein chains, are then treated as separate rigid bodies in a subsequent energy minimization stage. Using a biased Monte Carlo procedure, additional water molecules are removed from the interface: random water molecules are probed for their closest amino-acid / nucleotide residues on the partners and their probability to be kept is set equal to the observed fraction of water-mediated contacts for the specific water-mediated amino-acid nucleotide contact as obtained from the statistical analysis studies (Marabotti et al. 2008; Luscomb et al. 2001) described below. This procedure is repeated until 50% of the initial interfacial water molecules remain (*water_tokeep* parameter). Subsequently, water molecules with unfavorable interaction energies (sum of van der Waals and electrostatic water–protein energies > 0.0 kcal/mol) are removed. The latter procedure typically results in more than 50% of all waters removed from the interface. The best 20% of all solutions based on the HADDOCK score are then selected for further semi-flexible refinement in the follow up stages.

Semi-flexible simulated annealing (it1). For bound-bound docking, the molecules are kept rigid in this stage but the position of the water molecules is optimized. For bound-unbound docking the backbone and side-chain conformations of the protein interface(s) (within 5.0 Å of any partner molecule) are allowed to sample additional conformations in separate stages.

Water refinement (w). All solutions from it1 are subjected to a gentle refinement by solvating the complex in a primary layer of explicit water molecules.

Energetics and scoring. For protein-DNA docking we set the dielectric constant to

78.0 instead of the default 10.0 to damp the electrostatic contribution of DNA in vacuum. The overall HADDOCK score is calculated as a weighted sum of different terms, which are:

- for the rigid body stage: $0.01* E_{vdW} + 1.0 * E_{elec} + 0.01 * E_{AIR} – 0.01 * BSA + 1.0 * E_{desolv}$;
- for semi-flexible refinement: $1.0* E_{vdW} + 1.0 * E_{elec} + 0.1 * E_{AIR} – 0.01 * BSA + 1.0 * E_{desolv}$;
- and for the final water refinement: $1.0* E_{vdW} + 0.2 * E_{elec} + 0.1 * E_{AIR} + 1.0 * E_{desolv}$.

E_{vdW} is van der Waals energy, E_{elec} the electrostatic energy, E_{AIR} the ambiguous interaction restraints energy, BSA the buried surface area in $Å^2$ and E_{desolv} an empirical desolvation energy term (Fernández-Recio et al. 2004).

Protein-DNA water-mediated contact propensities

A statistical contact analysis of 100 high-resolution protein-DNA complexes performed by Marabotti *et al.* (Marabotti et al. 2008) yielded a propensity scale for water-mediated contacts between a given amino acid – base pair. Briefly, propensities were calculated by dividing the overall count of water-mediated contacts (reported in Table S7) to the overall count of amino acid-base interactions for each pair (reported in Table 2 of Marabotti et al. 2008). Additional propensities were derived for phosphate moieties from Luscombe *et al.* (Luscombe et al. 2001) in a similar manner (values are reported in Table 7 and Table 2 of Luscombe et al. 2001). Subsequently, obtained values were merged with the database of water-mediated amino acid –

amino acid contact propensities obtained from previous work (van Dijk and Bonvin 2006).

Analysis

Docking models were clustered based on their fraction of common contacts (Rodrigues et al. 2012) using a cut-off of 0.5 and minimum cluster size of 20. The best 20 solutions of the best cluster based on the HADDOCK score were selected and used for further analysis.

Model quality analysis. The quality of the generated solutions was evaluated using the CAPRI (Lensink and Wodak 2010b) criteria expressed as stars:

- three stars (high quality): Fnat > 0.5 and (l-RMSD or i-RMSD < 1.0 Å)
- two stars (medium quality): Fnat > 0.3 and (l-RMSD < 5.0 Å or i-RMSD < 2.0 Å)
- one star (acceptable quality): Fnat > 0.1 and (l-RMSD < 10.0 Å or i-RMSD < 4.0 Å).

Fnat is the fraction of native contacts within a 5.0 Å cutoff, i-RMSD is the interface backbone RMSD and l-RMSD is the ligand backbone RMSD calculated by superimposition on backbone atoms of the reference DNA (P,C1') and calculating the RMSDs on all backbone atoms of the reference protein ($C\alpha$,C,N,O) using an in-house fitting program based on fast quaternion-based methods (Liu et al. 2010; Theobald 2005).

Recovery analysis of interfacial waters. Water-mediated contact analysis was performed on all water molecules located within a 5.0 Å cut-off distance of the chains belonging to the protein-DNA interface using the NUCPLOT software package (Luscombe et al. 1997). A water-mediated contact was correctly reproduced if at least

one water-mediated contact was formed between any atom of the amino acid and any atom of the nucleotide using a 3.3Å cut-off distance (Schneider et al. 1992). A subdivision was made between contacts involving nucleotide base- and sugar-phosphate moieties. For the reference structure, only water molecules (identified by oxygen atoms) within a range defined by their average B-factor plus one times the standard deviation where considered. Interfacial water molecules where categorized as fully buried (defined by an accessible surface area of 0 $Å^2$ as measured by NACCESS (Hubbard and Thornton 1993)) or those that are positioned at the rim of the interface.

Acknowledgments

This work was supported by the Dutch Foundation for Scientific Research (NWO) through a VICI grant (n° 700.56.442) to A.M.J.J.B. and by the WeNMR project (European FP7 e-Infrastructure grant, contract no. 261572, www.wenmr.eu). The national Grid Initiatives of Belgium, France, Italy, Germany, The Netherlands (via the Dutch BiG Grid project), Portugal, Spain, U.K., South Africa, Taiwan, and the Latin America GRID infrastructure via the Gisela project are acknowledged for the use of computing and storage facilities. The European Grid Initiative (www.egi.eu) is acknowledged for its support of the WeNMR Virtual Research Community.

References

Ahmad M, Gu W, Geyer T, Helms V (2011) Adhesive water networks facilitate binding of protein interfaces. Nat Commun 2:261. doi: 10.1038/ncomms1258

Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7:188–197. doi: 10.1038/nrm1859

Baker D, Sali A (2001) Protein Structure Prediction and Structural Genomics. Science Signaling 294:93–96. doi: 10.1126/science.1065659

Ball P (2008) Water as an active constituent in cell biology. Chem Rev 108:74–108. doi: 10.1021/cr068037a

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301–3. doi: 10.1093/nar/gkl971

Brunger AT (2007) Version 1.2 of the Crystallography and NMR system. Nat Protoc 2:2728–2733. doi: 10.1038/nprot.2007.406

Chen L, Wang K, Shao Y, et al. (2008) Structural insight into the mechanisms of Wnt signaling antagonism by Dkk. J Biol Chem 283:23364–23370. doi: 10.1074/jbc.M802375200

Clore GM (2011) Exploring translocation of proteins on DNA by NMR. J Biomol NMR 51:209–219. doi: 10.1007/s10858-011-9555-8

Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. Nature 422:835–847. doi: doi:10.1038/nature01626

Collura V, Boissy G (2007) Subcellular Biochemistry. 43:135–183. doi: 10.1007/978-1-4020-5943-8_8

Cusick ME, Klitgord N, Vidal M, Hill DE (2005) Interactome: gateway into systems biology. Hum Mol Genet 14 Spec No. 2:R171–81. doi: 10.1093/hmg/ddi335

de Vries SJ, van Dijk ADJ, Krzeminski M, et al. (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins: Structure, Function, and Bioinformatics 69:726–733. doi: 10.1002/prot.21723

Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737. doi: 10.1021/ja026939x

Dunitz JD (1994) The entropic cost of bound water in crystals and biomolecules. Science 264:670. doi: 10.1126/science.264.5159.670

Dunn RK, Kingston RE (2007) Gene regulation in the postgenomic era: technology takes the wheel. Mol Cell 28:708–714. doi: 10.1016/j.molcel.2007.11.022

Fernández-Recio J, Totrov M, Abagyan R (2004) Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 335:843–865.

Giudice E, Lavery R (2002) Simulations of nucleic acids and their complexes. Acc Chem Res 35:350–357.

Horton NC (1998) Recognition of Flanking DNA Sequences by EcoRV Endonuclease Involves Alternative Patterns of Water-mediated Contacts. Journal of Biological Chemistry 273:21721–21729. doi: 10.1074/jbc.273.34.21721

Huang N, Shoichet BK (2008) Exploiting Ordered Waters in Molecular Docking. J

Med Chem 51:4862–4865. doi: 10.1021/jm8006239

Hubbard SJ, Thornton JM (1993) NACCESS.

Hummer G, García AE, Soumpasis DM (1995) Hydration of nucleic acid fragments: comparison of theory and experiment for high-resolution crystal structures of RNA, DNA, and DNA-drug complexes. Biophysj 68:1639–1652. doi: 10.1016/S0006-3495(95)80381-4

Janin J (1999) Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. Structure 7:R277–R279.

Jayaram B, Jain T (2004) The role of water in protein-DNA recognition. Annu Rev Biophys Biomol Struct 33:343–361. doi: 10.1146/annurev.biophys.33.110502.140414

Jones S, van Heyningen P, Berman HM, Thornton JM (1999) Protein-DNA interactions: A structural analysis. J Mol Biol 287:877–896. doi: 10.1006/jmbi.1999.2659

Joosten RP, Joosten K, Murshudov GN, Perrakis A (2012) PDB_REDO: constructive validation, more than just looking for errors. Acta Crystallogr D Biol Crystallogr 68:484–496. doi: 10.1107/S0907444911054515

Jorgensen WL, Chandrasekhar J, Madura JD (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935. doi: 10.1063/1.445869

Kalodimos CG, Biris N, Bonvin AMJJ, et al. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. Science 305:386–389. doi: 10.1126/science.1097064

Kastritis PL, van Dijk ADJ, Bonvin AMJJ (2012a) Explicit treatment of water molecules in data-driven protein-protein docking: the solvated HADDOCKing approach. Methods Mol Biol 819:355–374. doi: 10.1007/978-1-61779-465-0_22

Kastritis PL, Visscher KM, van Dijk ADJ, Bonvin AMJJ (2012b) Solvated protein-protein docking using Kyte-Doolittle-based water preferences. Proteins: Structure, Function, and Bioinformatics. doi: 10.1002/prot.24210

Lensink MF, Wodak SJ (2010a) Blind predictions of protein interfaces by docking calculations in CAPRI. Proteins: Structure, Function, and Bioinformatics 78:3085–3095. doi: 10.1002/prot.22850

Lensink MF, Wodak SJ (2010b) Docking and scoring protein interactions: CAPRI 2009. Proteins: Structure, Function, and Bioinformatics 78:3073–3084. doi: 10.1002/prot.22818

Li Z, Lazaridis T (2007) Water at biomolecular binding interfaces. Phys Chem Chem Phys 9:573–581. doi: 10.1039/b612449f

Liu LA, Bradley P (2012) Atomistic modeling of protein–DNA interaction

specificity: progress and applications. Curr Opin Struct Biol 22:397–405. doi: 10.1016/j.sbi.2012.06.002

Liu P, Agrafiotis DK, Theobald DL (2010) Fast determination of the optimal rotational matrix for macromolecular superpositions. J Comput Chem 1561–1563. doi: 10.1002/jcc.21439

Luscombe NM, Laskowski RA, Thornton JM (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res 29:2860-2874.

Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol 1:1–10. doi: 10.1186/gb-2000-1-1-reviews001

Luscombe NM, Laskowski RA, Thornton JM (1997) NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. Nucleic Acids Res 25:4940–4945. doi: 10.1093/nar/29.13.2860

Makarov V, Pettitt BM, Feig M (2002) Solvation and hydration of proteins and nucleic acids: a theoretical view of simulation and experiment. Acc Chem Res 35:376–384.

Marabotti A, Spyrakis F, Facchiano A, et al. (2008) Energy-based prediction of amino acid-nucleotide base recognition. J Comput Chem 29:1955–1969. doi: 10.1002/jcc.20954

Melquiond AS, Karaca E, Kastritis PL, Bonvin AM (2011) Next challenges in protein-protein docking: from proteome to interactome and beyond. WIREs Comput Mol Sci 642–651. doi: 10.1002/wcms.91

Moitessier N, Westhof E, Hanessian S (2006) Docking of aminoglycosides to hydrated and flexible RNA. J Med Chem 49:1023–1033. doi: 10.1021/jm0508437

Nadassy K, Wodak SJ, Janin J (1999) Structural features of protein-nucleic acid recognition sites. Biochemistry 38:1999–2017. doi: 10.1021/bi982362d

Nederveen AJ, Doreleijers JF, Vranken W, et al. (2005) RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins: Structure, Function, and Bioinformatics 59:662–672. doi: 10.1002/prot.20408

Otwinowski Z, Schevitz RW, Zhang RG, et al. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. Nature 335:321–329. doi: 10.1038/335321a0

Pakotiprapha D, Jeruzalmi D (2009) Crystallization of Protein-DNA Complexes. Encyclopedia of Life Sciences (ELS). doi: 10.1002/9780470015902.a0002720.pub2

Pandey A, Mann M (2000) Proteomics to study genes and genomes. Nature 405:837–846. doi: 10.1038/35015709

Pérez A, Luque FJ, Orozco M (2012) Frontiers in molecular dynamics simulations of DNA. Acc Chem Res 45:196–205. doi: 10.1021/ar2001217

Reddy CK, Das A, Jayaram B (2001) Do water molecules mediate protein-DNA recognition? J Mol Biol 314:619–632. doi: 10.1006/jmbi.2001.5154

Renuse S, Chaerkady R, Pandey A (2011) Proteogenomics. Proteomics 11:620–630. doi: 10.1002/pmic.201000615

Roberts VA, Case DA, Tsui V (2004) Predicting interactions of winged-helix transcription factors with DNA. Proteins: Structure, Function, and Bioinformatics 57:172–187. doi: 10.1002/prot.20193

Rodrigues JPGLM, Trellet M, Schmitz C, et al. (2012) Clustering biomolecular complexes by residue contacts similarity. Proteins: Structure, Function, and Bioinformatics 80:1810–1817. doi: 10.1002/prot.24078

Samish I (2009) Search and sampling in structural bioinformatics. In: Gu J, Bourne PE (eds) Structural bioinformatics. John Wiley & Sons, Hoboken, New Jersey, pp 207–235

Schneider B, Cohen D, Berman HM (1992) Hydration of DNA bases: Analysis of crystallographic data. Biopolymers 32:725–750. doi: 10.1002/bip.360320703

Schwabe JW (1997) The role of water in protein-DNA interactions. Curr Opin Struct Biol 7:126–134.

Shakked Z, Guzikevich-Guerstein G, Frolow F, et al. (1994) Determinants of repressor/operator recognition from the structure of the trp operator binding site. Nature 368:469–473. doi: 10.1038/368469a0

Sidorova NY, Rau DC (1996) Differences in water release for the binding of EcoRI to specific and nonspecific DNA sequences. Proc Natl Acad Sci USA 93:12272–12277.

Spyrakis F, Cozzini P, Bertoli C, et al. (2007) Energetics of the protein-DNA-water interaction. BMC Struct Biol 7:4. doi: 10.1186/1472-6807-7-4

Theobald DL (2005) Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. Acta Crystallogr, A, Found Crystallogr 61:478–480. doi: 10.1107/S0108767305015266

van Dijk ADJ, Boelens R, Bonvin AMJJ (2005) Data-driven docking for the study of biomolecular complexes. FEBS J 272:293–312. doi: 10.1111/j.1742-4658.2004.04473.x

van Dijk ADJ, Bonvin AMJJ (2006) Solvated docking: introducing water into the modelling of biomolecular complexes. Bioinformatics 22:2340–2347. doi: 10.1093/bioinformatics/btl395

van Dijk M, Bonvin AMJJ (2010) Pushing the limits of what is achievable in protein-DNA docking: benchmarking HADDOCK's performance. Nucleic Acids Res

38:5634–5647. doi: 10.1093/nar/gkq222

van Dijk M, Bonvin AMJJ (2008) A protein-DNA docking benchmark. Nucleic Acids Res 36:e88. doi: 10.1093/nar/gkn386

van Dijk M, Bonvin AMJJ (2009) 3D-DART: a DNA structure modelling server. Nucleic Acids Res 37:W235–9. doi: 10.1093/nar/gkp287

van Dijk M, van Dijk ADJ, Hsu V, et al. (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. Nucleic Acids Res 34:3317–3325. doi: 10.1093/nar/gkl412

Varani G, Chen Y, Leeper TC (2004) NMR studies of protein-nucleic acid interactions. Methods Mol Biol 278:289–312. doi: 10.1385/1-59259-809-9:289

Virtanen JJ, Makowski L, Sosnick TR, Freed KF (2010) Modeling the Hydration Layer around Proteins: HyPred. Biophys J 99:1611–1619. doi: 10.1016/j.bpj.2010.06.027

Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics 38:95–109. doi: 10.1016/j.jgg.2011.02.003

Table 1. The solvated protein-DNA benchmark

Complex		Protein		Water molecules				
PDB id[a]	**Cat.**[b]	**PDB id**[c]	**Description**	**Total**[d]	**Buried**[e]	**Nr.**[f]	**BSA**[g]	**Inter.**[h]
1a73	8	1evx^{X}	Intron-encoded homing endonuclease I-PPOI	74	46	2	18344	1.19
1azp	6	1sap^{X}	Hyperthermophile chromosomal protein SAC7D	10	7	2	6533	2.79
1bgb	8	1rve^{X}	Endonuclease EcoRV	47	21	2	20033	3.92
1bnz	6	1sso^{X}	Hyperthermophile chromosomal protein SSO7D	8	5	2	6107	2.77
1cdw	5	1vok^{X}	Human TATA-box binding protein core domain	18	4	2	12574	0.80
1ckq	8	1qc9^{X}	Endonuclease EcoRI pre-transition state	30	24	3	23558	2.09
1eyu	8	1pvu^{X}	Endonuclease PvuII	82	66	2	15661	6.34
1fjl	2	3a02^{X}	Drosophila paired protein homeodomain	39	30	3	9833	0.56
1g9z	8	2o7m^{X}	Homing endonuclease I-CreI	70	55	2	16678	4.21
1mnn	1	1mn4^{X}	Sporulation specific transcription factor Ndt80	28	24	2	17486	0.77
1qne	5	1vok^{X}	Arabidopsis transcription initiation factor TFIID-1	14	5	2	12594	0.89
1rh6	8	1lx8^{N}	Bacteriophage Lambda excisionase (Xis)	17	4	2	11318	1.40 $^{0.50}$
1rva	8	1rve^{X}	Endonuclease EcoRV	59	43	2	21800	3.87

1w0t	1	1ba5^{N}	Human telomeric repeat binding factor 1	18	19	3	11728	1.75 $^{0.45}$
1ytb	5	1tbp^{X}	S. cerevisiae TATA-box binding protein	18	10	2	12891	1.15
1zs4	1	1zpq^{X}	Bacteriophage lambda regulator protein cII	27	3	2	24171	3.78
1ztw	8	1mml^{X}	Moloney murine leukemia virus reverse transcriptase catalytic fragment	5	1	2	15513	1.82
2hdd	2	1enh^{X}	Drosophila engrailed homeodomain	29	19	3	13382	0.72
2itl	4	1tbd^{N}	Simian virus 40 large T antigen origin-binding domain	15	1	3	18949	1.00 $^{0.12}$
2o4a	1	1yse^{N}	Transcription factor SATB1 CUT domain	9	2	2	8881	1.95 $^{0.69}$
2oaa	8	2oa9^{X}	Restriction endonuclease MvaI	83	81	2	12518	7.96
2odi	8	2odh^{X}	Restriction endonuclease BcnI	64	59	2	12706	3.14
2r1j	1	1adr^{N}	P22 c2 repressor protein	18	17	3	12098	1.06 $^{0.24}$
3bam	8	1bam^{X}	Restriction endonuclease BamHI	52	40	3	17947	5.03
3jxy	8	3bvs^{X}	B. cereus alkylpurine DNA glycosylase AlkD	19	3	2	14645	0.48
3kxt	4	2jtm^{N}	S. solfataricus chromatin protein Cren7	14	0	2	5810	1.15 $^{0.67}$
3ted	6	2xb0^{X}	S. cerevisiae chromo domain-containing protein 1	24	2	2	16827	2.15
3v6t	4	3v6p^{X}	Xanthomonas dHax3 TAL-effector protein	66	59	2	23740	6.25

(**a**) RCSB PDB accession number for the structure of the complex and the unbound protein (**c**). Structures for the unbound protein were either solved by X-ray crystallography^{X} or NMR spectroscopy^{N}.

(**b**) Classification of the protein–DNA complexes in 6 out of eight different groups (Luscombe et al. 2000); helix–turn–helix (1), zinc-coordinating (2), other a-helix (4), b-sheet (5), b-hairpin/ribbon 6) and enzyme (8) complexes.

(**d**) Total number of interfacial water molecules within 5 Å of any fully buried protein-DNA interface residue (amino-acid or nucleotide).

(**e**) Total number of fully buried interfacial water molecules as defined by an accessible surface area of 0 $Å^2$ as calculated by NACCESS. Only water molecules within a range defined by their average B-factor plus one times the standard deviation where considered.

(**f**) Number of individual biomolecules that need to be docked to reconstruct the complex.

(**g**) Buried surface area of the DNA upon complex formation in $Å^2$.

(**h**) RMSD (Å) from the bound form of the protein calculated over the interface atoms ($C\alpha$, C, N, O) of the unbound protein structure after superposition onto the reference complex.

Table 2. Interface RMSD, fraction of native contacts and interface water recovery for the best 20 docking solutions in bound-bound and bound-unbound, solvated and non-solvated docking.

	Bound-Bound docking							Bound-Unbound docking					
PDB	**Non-solvated**			**Solvated**				**Non-solvated**			**Solvated**		
id[a]	**iRMSD[b]**	**fNAT[c]**	*	**iRMSD[b]**	**fNAT[c]**	**H_2O recov[d]**	*	**iRMSD[b]**	**fNAT[c]**	*	**iRMSD[b]**	**fNAT[c]**	*
3jxy	0.53 $_{0.04}$	0.59 $_{0.02}$	3	0.55 $_{0.04}$	0.60 $_{0.01}$	0.33 $_{0.18}$	3	1.21 $_{0.26}$	0.40 $_{0.08}$	2	1.48 $_{0.19}$	0.28 $_{0.05}$	2
1fjl	0.62 $_{0.04}$	0.66 $_{0.01}$	3	0.57 $_{0.05}$	0.69 $_{0.02}$	0.10 $_{0.09}$	3	1.23 $_{0.21}$	0.06 $_{0.01}$	0	1.09 $_{0.17}$	0.07 $_{0.01}$	0
1rxw	0.43 $_{0.03}$	0.71 $_{0.02}$	3	0.40 $_{0.03}$	0.73 $_{0.02}$	0.07 $_{0.07}$	3	2.76 $_{0.05}$	0.38 $_{0.06}$	0	2.68 $_{0.09}$	0.41 $_{0.03}$	0
2hdd	0.66 $_{0.09}$	0.61 $_{0.03}$	3	0.58 $_{0.05}$	0.61 $_{0.02}$	0.11 $_{0.12}$	3	3.64 $_{0.29}$	0.17 $_{0.03}$	0	2.07 $_{0.37}$	0.22 $_{0.04}$	0
1mnn	0.65 $_{0.05}$	0.64 $_{0.02}$	3	0.72 $_{0.07}$	0.59 $_{0.03}$	0.19 $_{0.09}$	2	0.90 $_{0.10}$	0.44 $_{0.03}$	2	0.98 $_{0.07}$	0.42 $_{0.03}$	2
1cdw	0.47 $_{0.03}$	0.80 $_{0.02}$	3	0.48 $_{0.04}$	0.79 $_{0.02}$	0.24 $_{0.11}$	3	1.31 $_{0.31}$	0.38 $_{0.11}$	2	1.05 $_{0.06}$	0.47 $_{0.04}$	2
1qne	0.41 $_{0.02}$	0.83 $_{0.01}$	3	0.40 $_{0.02}$	0.83 $_{0.01}$	0.10 $_{0.06}$	3	0.85 $_{0.12}$	0.58 $_{0.04}$	3	1.09 $_{0.17}$	0.46 $_{0.06}$	2
2itl	0.57 $_{0.09}$	0.64 $_{0.01}$	3	0.55 $_{0.05}$	0.65 $_{0.01}$	0.11 $_{0.05}$	3	2.51 $_{0.58}$	0.15 $_{0.07}$	0	1.77 $_{0.34}$	0.18 $_{0.05}$	0
2r1j	0.38 $_{0.06}$	0.65 $_{0.01}$	3	0.41 $_{0.05}$	0.65 $_{0.01}$	0.24 $_{0.06}$	3	1.23 $_{0.66}$	0.39 $_{0.10}$	2	1.19 $_{0.25}$	0.46 $_{0.04}$	2
1ytb	0.44 $_{0.02}$	0.84 $_{0.01}$	3	0.20 $_{0.03}$	0.81 $_{0.03}$	0.23 $_{0.09}$	3	0.99 $_{0.12}$	0.54 $_{0.05}$	2	1.00 $_{0.10}$	0.52 $_{0.04}$	2
3kxt	0.48 $_{0.08}$	0.80 $_{0.03}$	3	0.47 $_{0.04}$	0.80 $_{0.03}$	0.31 $_{0.10}$	3	1.38 $_{0.16}$	0.42 $_{0.08}$	2	1.59 $_{0.17}$	0.37 $_{0.07}$	2

1a73	0.59 $_{0.02}$	0.62 $_{0.02}$	3	0.54 $_{0.02}$	0.63 $_{0.02}$	0.40 $_{0.05}$	3	1.37 $_{0.15}$	0.39 $_{0.03}$	2	1.42 $_{0.10}$	0.37 $_{0.03}$	2
1rh6	0.62 $_{0.04}$	0.58 $_{0.03}$	2	0.65 $_{0.07}$	0.70 $_{0.02}$	0.04 $_{0.09}$	2	2.66 $_{0.25}$	0.23 $_{0.06}$	0	2.36 $_{0.41}$	0.24 $_{0.05}$	0
1tro	0.61 $_{0.03}$	0.72 $_{0.01}$	3	1.14 $_{0.32}$	0.61 $_{0.07}$	0.10 $_{0.06}$	3	2.61 $_{0.39}$	0.07 $_{0.02}$	0	3.55 $_{0.31}$	0.03 $_{0.01}$	0
1w0t	0.55 $_{0.06}$	0.59 $_{0.02}$	3	0.56 $_{0.04}$	0.58 $_{0.03}$	0.11 $_{0.05}$	3	2.34 $_{0.42}$	0.27 $_{0.04}$	2	2.25 $_{0.35}$	0.26 $_{0.04}$	2
1ztw	0.95 $_{0.31}$	0.68 $_{0.08}$	3	0.75 $_{0.15}$	0.71 $_{0.07}$	0.00 $_{0.00}$	2	1.03 $_{0.35}$	0.41 $_{0.08}$	2	1.35 $_{0.39}$	0.35 $_{0.11}$	2
2o4a	0.63 $_{0.08}$	0.60 $_{0.02}$	3	0.50 $_{0.05}$	0.66 $_{0.02}$	0.12 $_{0.09}$	3	1.96 $_{0.69}$	0.32 $_{0.09}$	2	1.35 $_{0.51}$	0.37 $_{0.08}$	2
1ckq	0.46 $_{0.04}$	0.74 $_{0.01}$	3	0.49 $_{0.03}$	0.74 $_{0.01}$	0.06 $_{0.07}$	3	5.66 $_{0.15}$	0.29 $_{0.01}$	0	5.71 $_{0.18}$	0.29 $_{0.01}$	0
3ted	0.89 $_{0.29}$	0.47 $_{0.05}$	3	0.72 $_{0.06}$	0.49 $_{0.02}$	0.16 $_{0.13}$	3	2.18 $_{0.14}$	0.28 $_{0.03}$	0	2.01 $_{0.17}$	0.29 $_{0.02}$	0
1bnz	0.57 $_{0.05}$	0.76 $_{0.01}$	3	0.55 $_{0.05}$	0.75 $_{0.02}$	0.22 $_{0.11}$	3	2.30 $_{0.29}$	0.41 $_{0.06}$	2	2.70 $_{0.27}$	0.33 $_{0.08}$	0
1azp	0.51 $_{0.04}$	0.79 $_{0.02}$	3	0.45 $_{0.04}$	0.78 $_{0.03}$	0.19 $_{0.12}$	3	2.04 $_{0.18}$	0.23 $_{0.05}$	2	2.21 $_{0.09}$	0.18 $_{0.04}$	0
1zs4	0.45 $_{0.02}$	0.61 $_{0.01}$	3	0.42 $_{0.03}$	0.62 $_{0.01}$	0.40 $_{0.12}$	3	5.19 $_{0.20}$	0.05 $_{0.02}$	0	5.57 $_{0.11}$	0.02 $_{0.01}$	0
1g9z	0.63 $_{0.05}$	0.58 $_{0.03}$	3	0.53 $_{0.03}$	0.62 $_{0.01}$	0.12 $_{0.05}$	3	2.83 $_{0.16}$	0.23 $_{0.03}$	0	3.36 $_{0.18}$	0.12 $_{0.02}$	0
3bam	1.38 $_{0.30}$	0.22 $_{0.04}$	2	0.85 $_{0.34}$	0.60 $_{0.10}$	0.17 $_{0.11}$	3	8.89 $_{0.29}$	0.18 $_{0.01}$	0	7.93 $_{0.34}$	0.23 $_{0.01}$	0

(**a**) RCSB PDB accession number for the structure of the complex sorted from low to high backbone RMS deviations between bound and unbound protein conformations.

(b) Average and standard deviation for interface RMSD values calculated by superimposition of all backbone atoms (C α , C, N, O, P and C1') of the interface residues with respect to the target and fNAT

(c) Fraction of native contacts within a 5.0 Å cutoff.

(d) Average and standard deviation for the fraction of recovered interfacial water molecules that are within 1.5Å for the target water oxygen atom.

(*) The CAPRI quality score expressed as the number of stars of the top ranking solution; not acceptable (0), acceptable (1), intermediate (2) or high quality (3).

The significance in the difference of means between solvated and non-solvated docking runs was calculated using a two-sample t-test with a symmetric 95% confidence interval. A significant improvement in performance is indicated in light gray and a decrease in dark gray boxes.

Table 3. Performance of the solvated docking protocol with respect to iRMSD and fNAT of the 10 best models when applied to our two-stage protein-DNA docking protocol using unbound partners and knowledge based AIRs.

	Unbound-Unbound docking			
	Non-solvated		**Solvated[d]**	
PDB id[a]	**iRMSD[b]**	**fNAT[c]**	**iRMSD[b]**	**fNAT[c]**
'Easy'				
1by4	4.91 $_{2.32}$	0.27 $_{0.09}$	3.04 $_{0.16}$	0.29 $_{0.02}$
3cro	2.62 $_{0.75}$	0.40 $_{0.06}$	2.17 $_{0.18}$	0.41 $_{0.02}$
'Intermediate'				
1azp	4.00 $_{0.45}$	0.10 $_{0.04}$	3.60 $_{0.57}$	0.14 $_{0.01}$
1jj4	3.62 $_{0.38}$	0.21 $_{0.07}$	3.38 $_{0.18}$	0.22 $_{0.03}$
'Difficult'				
1a74	$3.37_{0.32}$	$0.24_{0.05}$	$3.37_{0.15}$	$0.15_{0.02}$
1zme	$4.63_{0.80}$	$0.15_{0.04}$	$5.22_{0.05}$	$0.14_{0.01}$

(**a**) RCSB PDB accession number for the structure of the complex sorted according to their difficulty as defined in (van Dijk and Bonvin 2008).

(**b**) Interface RMSD values calculated by superimposition of all backbone atoms (C α , C, N, O, P and C1') of the interface residues with respect to the target.

(**c**) Fraction of native contacts within a 5.0 Å cutoff.

(**d**) The significance in the difference of means between solvated and non-solvated docking runs was calculated using a two-sample t-test with a symmetric 95% confidence interval. A significant improvement in performance is indicated in light gray and a decrease in dark gray boxes.

Figure 1: Correlation plot between the average total (blue) and fully buried (red) number of observed interfacial water molecules in the 20 best solutions with respect to the reference complexes for bound-bound **(a)** and bound-unbound **(b)** solvated docking. The 20 best models are selected based on the HADDOCK score after water refinement and clustering. Interfacial water molecules are calculated as those water molecules within 5.0 Å of any fully buried interface residue thus including water molecules located at the rim region of the interface, fully buried water molecules are defined as those with zero surface accessibility. The Pearson's correlation coefficients and confidence intervals between modeled- and crystal interface water molecules are shown as inset in the figures. Visual overlay of a representative bound-bound docking model (blue) and the reference complex (red) for **(c)** the Intron-encoded homing endonuclease I-PPOI (1a73 side view) and **(d)** the TATA-box binding protein (**d**, top view). Oxygen atoms of the water molecules are shown as spheres.

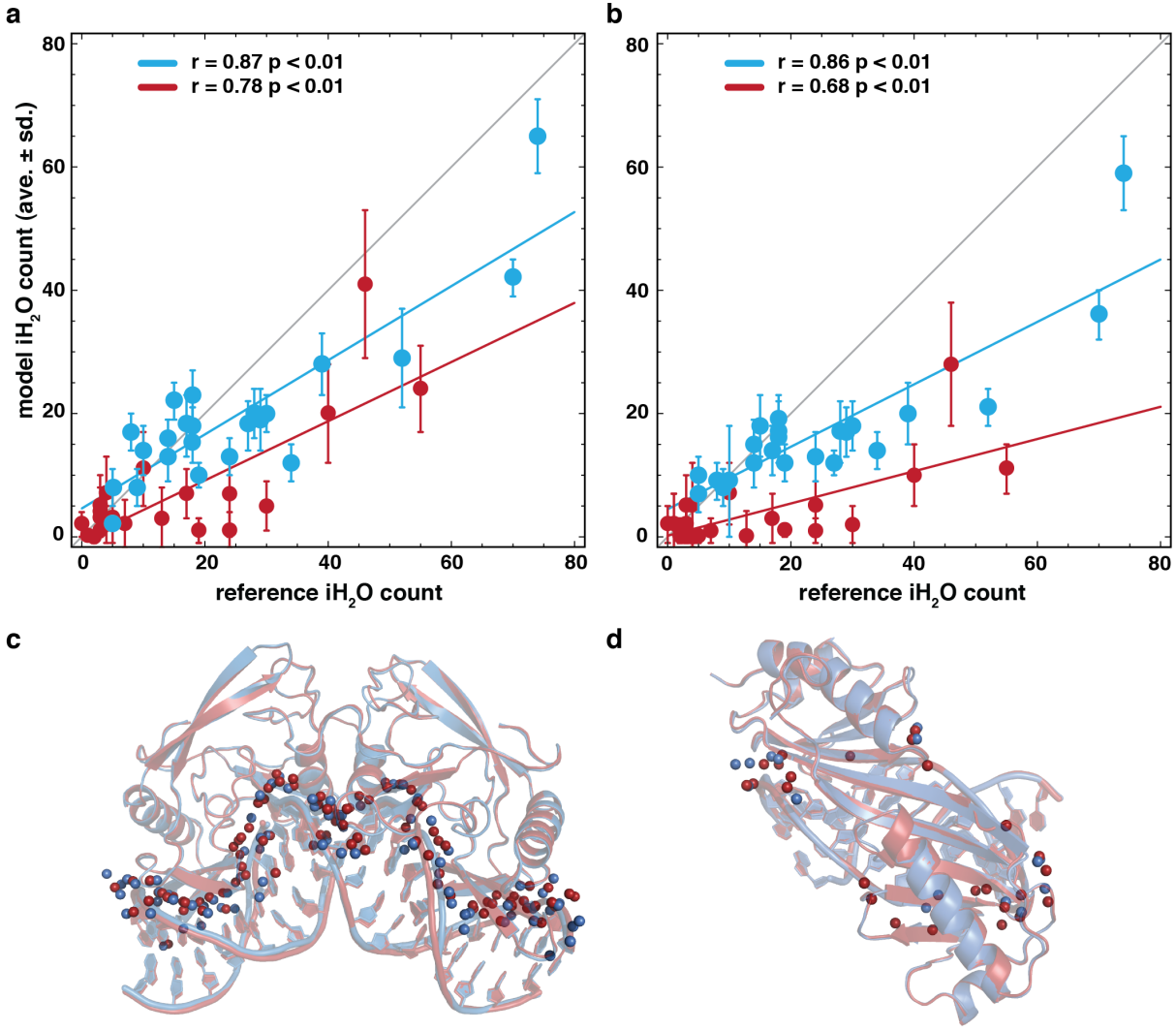
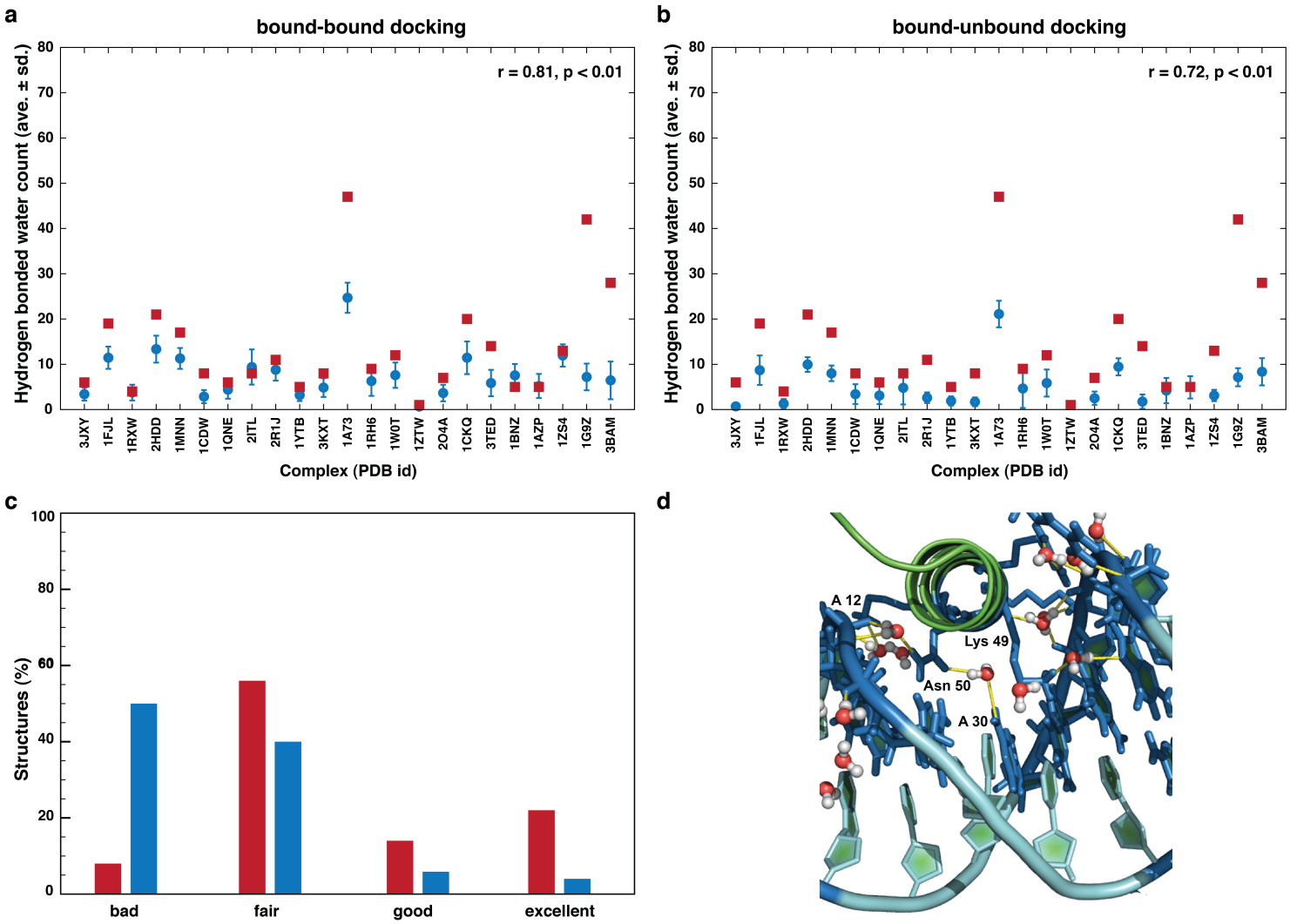

a
b
c
d
r = 0.87 p < 0.01
r = 0.78 p < 0.01
r = 0.86 p < 0.01
r = 0.68 p < 0.01
model iH2O count (ave. ± sd.)
reference iH2O count
reference iH2O count
0
20
40
60
80

Figure 2: Overall success rate for water-mediated contact recovery. Average and standard deviations for interfacial hydrogen-bonded water count (blue circles ●) in the 20 best bound-bound (**a**) and bound-unbound (**b**) solvated docking solutions after water refinement and clustering shown together with their corresponding reference target value (red squares ■). Hydrogen-bonded interfacial water molecules where identified using the NUCPLOT (Luscombe et al. 1997) software package. The benchmark cases are sorted based on the amount of conformational changes, from low to high backbone RMS deviations between bound and unbound protein conformations. The Pearson's correlation coefficients and confidence intervals between modeled- and crystal hydrogen-bonded interface water molecules are shown as inset in the figures. **(c)** Percentage of the 20 best solutions of all benchmark cases with the corresponding fraction of native water-mediated contacts CAPRI quality score for bound-bound (red squares ■) and bound-unbound (blue circles ●) docking classified as: bad = $f^{w}_{nat} < 0.1$, fair = $0.1 \leq f^{w}_{nat} < 0.3$, good = $0.3 \leq f^{w}_{nat} < 0.5$, excellent = $0.5 \leq f^{w}_{nat} < 0.8$. **(d)** Specific water-mediated contact recovery as observed in 50% of the selected docking solutions of the Drosophila engrailed homeodomain (PDB id: 2hdd) viewed along the DNA recognition helix positioned in the DNA major groove. Residue labels in the reference structure indicate the correctly predicted contacts.


a
bound-bound docking
r = 0.81, p < 0.01
Hydrogen bonded water count (ave. ± sd.)
Complex (PDB id)
3JXY 1FJL 1RXW 2HDD 1MNN 1CDW 1QNE 2ITL 2R1J 1YTB 3KXT 1A73 1RH6 1W0T 1ZTW 2O4A 1CKQ 3TED 1BNZ 1AZP 1ZS4 1G9Z 3BAM
b
bound-unbound docking
r = 0.72, p < 0.01
Hydrogen bonded water count (ave. ± sd.)
Complex (PDB id)
3JXY 1FJL 1RXW 2HDD 1MNN 1CDW 1QNE 2ITL 2R1J 1YTB 3KXT 1A73 1RH6 1W0T 1ZTW 2O4A 1CKQ 3TED 1BNZ 1AZP 1ZS4 1G9Z 3BAM
c
Structures (%)
bad
fair
good
excellent
d
A 12
Lys 49
Asn 50
A 30