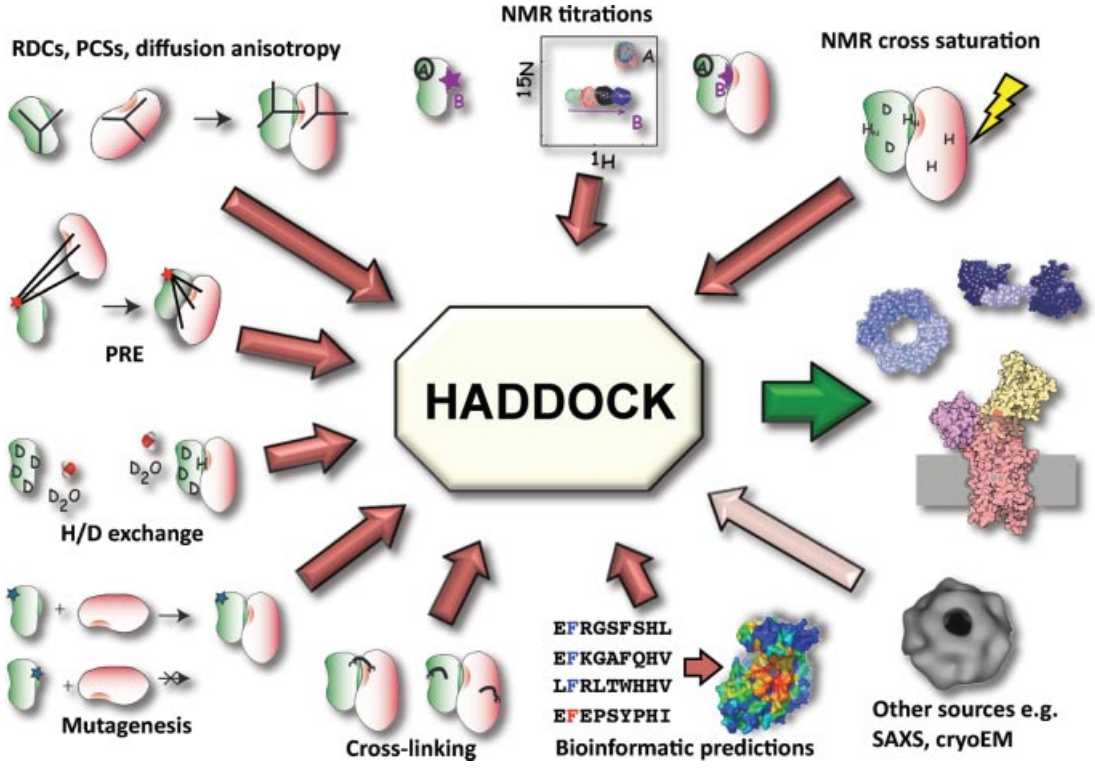
Information sources for data-driven HADDOCKing



HADDOCK can use a variety of experimental information during the docking of protein complexes. Initially developed to exploit chemical shift perturbation data (measured by NMR titrations), HADDOCK slowly "learned" to use more and more NMR and non-NMR information. Future development should include the possibility to use SAXS and cryo-electron microscopy data.

32
Protein–Protein Docking with HADDOCK

Christophe Schmitz, Adrien S.J. Melquiond, Sjoerd J. de Vries, Ezgi Karaca, Marc van Dijk, Panagiotis L. Kastritis, and Alexandre M.J.J. Bonvin

Advances in biophysics and biochemistry have pushed back the limits of the structural characterization of biomolecular assemblies. Mixing even a limited amount of experimental and/or bioinformatics data with modeling methods such as macromolecular docking represents a valuable strategy to predict the three-dimensional structures of complexes. In this chapter, we discuss the HADDOCK data-driven approach to the modeling of complexes. The program supports a wide range of NMR and other experimental data as well as bioinformatics predictions. It is also available as a user-friendly web server, facilitating the modeling of biomolecular complexes for a wide community.

32.1
Protein–Protein Docking: General Concepts

32.1.1
Why Protein–Protein Docking?

Techniques to obtain the atomic structures of single proteins are maturing. The size limit of macromolecules that we can accurately model is continually expanding and the number of structures deposited in the Protein Data Bank (PDB) has followed a nearly exponential growth over the past 20 years, reaching 70 231 entries in early 2011.[1)]

One of the remaining challenges, however, is obtaining assemblies of two or more macromolecules (proteins, DNA, RNA). On average, it is believed that each protein interacts with about 8–10 other macromolecules, but only 2990 complexes are present in the PDB. For X-ray crystallography, the difficulty in cocrystallizing a complex is much greater than for individual proteins. For NMR spectroscopy, the large molecular weight of complexes presents a problem, making it more difficult to obtain and analyze data. Furthermore, intermolecular nuclear Overhauser effects (NOEs), which provide the most useful information, often involve amino acid side-chains, the resonances of which are much harder to assign than the backbone resonances of the protein. Currently, the only method that can systematically give insights into the large number of protein complexes encountered in biological processes is protein docking *in silico*. This consists of predicting the binding mode of protein complexes starting from their *free-form (unbound)* experimental or modeled individual three-dimensional structure. Several software packages have been developed for this purpose. The majority of them try to predict protein–protein complexes using solely geometrical and/or energetic considerations [1,2]. HADDOCK [3–5] distinguishes itself by including experimental, notably NMR, data and/or bioinformatics information to efficiently drive the docking process.

1) http://www.rcsb.org/pdb/statistics/holdings.do.

NMR of Biomolecules: Towards Mechanistic Systems Biology, First Edition. Edited by Ivano Bertini, Kathleen S. McGreevy, and Giacomo Parigi.

32.1.2 General Methods for Protein–Protein Docking

To achieve efficient protein–protein docking, the most common strategy is to combine an efficient sampling of a large number of possible conformations of the complex with an accurate scoring of each of them, in order to devise a model similar to the native one. Common sampling strategies include Monte Carlo minimization (e.g., as used in RosettaDock [6] and ICM-DISCO [7]) and fast Fourier transform algorithms (e.g., as used in ZDOCK [8], MolFit [9], CLUSPRO [10], DOT [11], GRAMM-X [12], and PIPER [13]). The incorporation of molecular dynamics steps is usually reserved for optimization in the final stages. The scoring (i.e., the calculation of some energy or score function for a given complex) usually consists of a combination of several biophysical and/or empirical energy terms, such as van der Waals energy, electrostatic energy, desolvation energy, buried surface area, geometric surface matching, and so on. Various scoring schemes can be used depending on the level of complexity, with more simple functions often used in the initial stage of the search (where a very large number of conformations are sampled) and more sophisticated (and costly in terms of computing time) used in the later refinement stages. Several hundreds, thousands, or tens of thousands of models are usually generated, depending on the docking approach. Clustering of these models is often performed to select the best predictions. All docking software platforms try to find the most efficient combination of sampling method and scoring function in order to derive the most accurate complex in the shortest CPU time.

Recent results in the CAPRI challenge, which aimed to assess the quality of various docking software with blind predictions [14], indicate that most methods can accurately predict the complex of proteins when their separate bound forms are provided [15]. However, bound–bound docking is only an artificial test to assess the performance of docking methods, the real challenge being to predict the structure of a complex from its free, unbound constituents [16]. Crystal or NMR structures can be used as starting structures when available and otherwise homology models when possible. CAPRI results have revealed that small conformational changes (i.e., around 2 Å backbone root mean square deviation (RMSD)) between the bound and free forms already present a significant challenge, and that with larger conformational changes, the sampling of near native complexes often becomes impossible [17,18]. To overcome those difficulties, some docking software platforms allow flexibility of the side-chains and/or backbone.

HADDOCK is one of the few docking software platforms that explicitly takes flexibility into account both in the side-chains and backbone of the proteins. In addition, large conformational changes can be modeled by considering ensembles of starting conformations or even by treating molecules as a collection of subdomains. To focus the sampling of conformations around the relevant interfaces, HADDOCK proposes a hybrid method that can include a large variety of NMR (and non-NMR) experimental information to guide the docking process. The main idea is to gather easily obtained information in order to drive the docking process and improve the scoring of the generated models.

32.2 Gathering Experimental Information for Data-Driven Docking

Along with other docking software, HADDOCK has an *ab initio* mode that allows the docking of macromolecules in the absence of any experimental data. However, to increase the chances of a successful docking, incorporation of even a limited amount of experimental data has proven to be valuable. In this section, we describe all the classes of NMR experiments (see also Section 9.4.3) that provide useful intermolecular information that can be used by HADDOCK. For completeness, the Section 32.2.9 will give an overview of non-NMR methods that can (or will soon) also be handled by HADDOCK. Table 32.1 summarizes the pro and cons of each technique.

32.2.1 Chemical Shift Perturbations

Chemical shift perturbations (CSPs) [19] are an easy way to gain information about the residues involved at the interface of a complex. As already described in Section 9.4.3.1, the experiment consists of tracking the chemical shift displacement of an NMR spectrum upon titration of another (unlabeled) molecule. Typically, a ^{15}N heteronuclear single-quantum coherence (HSQC) spectrum is first recorded with only one of the protein partners. Subsequently, as the concentration of the second protein is slowly increased, the recorded spectra are expected to exhibit changes due to the formation of the complex. The chemical shifts of the spin experiencing local environment changes will be displaced. This phenomenon mainly occurs at the interface of the protein upon backbone and/or side-chain rearrangements, upon electronic interaction with the other protein partner, or solvent reshuffling at the interface. It cannot be excluded, however, that residues located far away from the interface also experience conformational changes (allosteric effects) and will hence appear as false positives. False negatives can also be encountered as nothing guarantees that the chemical shifts of

Table 32.1 List of experimental data that HADDOCK supports (or soon will), together with the advantages and disadvantages of each type of data, and some remarks.

	Experimental data	Outcome	Advantages	Drawbacks	Remarks
NMR	CSP	identification of residue located at the interface	easily conducted	false positives, false negatives	
	cross-saturation experiments	identification of residue located at the interface	accurate identification of the interface	requires deuteration of one the protein	
	hydrogen/deuterium exchange	identification of residue located at the interface	accurate identification of the interface	false positives, false negatives	can also be monitored with MS
	NOE	proton distance information	effective restraint	intermolecular NOE difficult to obtain	spin diffusion can induce false positives
	PRE	distance information from the paramagnetic ion		requires paramagnetic labeled protein, usually with a paramagnetic tag	paramagnetic center needs to be close to the interface to observe effect
	PCS	distance and angular information from the paramagnetic ion	long-range distance, multiple independent datasets	requires paramagnetic labeled protein, usually with a paramagnetic tag	
	RDC	orientational information	multiple independent datasets	requires alignment media (can also by aligned with a paramagnetic center)	
	diffusion anisotropy	orientational information			type of restraint similar to RDC
Other	interface prediction	prediction of residue located at the interface	no experiments required	quality of prediction subject to presence of homologous in databases	efficient for obligate complexes
	site-directed mutagenesis	identification of residue located at the interface		false positives, false negatives	
	cryo-electron microscopy	overall surface/shape of the complex	overall shape of large assembly; size larger than 110 kDa	low-resolution information. risk of noise contamination	not yet implemented as restraint in HADDOCK, but can be used in scoring
	SAXS	overall surface/shape of the complex	fast; small concentration of protein required; size 50–250 kDa	low-resolution information	not yet implemented as restraint in HADDOCK, but can be used in scoring
	cross-linking	upper distance between cross-links		bounded distance can be large, cross-links can disturb the native state	some cross-links are nonspecific leading to risk of false positives

spins located at the interface do change. It is therefore important to carefully interpret the CSP in order to extract the residues involved in the complex formation process. Mapping the interface of the second protein is done by repeating the same protocol with the role inversed (first protein titrated, second protein labeled). Later in this chapter, we will see how HADDOCK can exploit this interface information.

32.2.2 Cross-Saturation Experiments

In cross-saturation experiments [20], one of the two proteins is $^{2}H/^{15}N$ uniformly labeled, containing only those protons that can be exchanged as the observable protons (e.g., the amide protons). The applied radiofrequency field will only irradiate the unlabeled protein protons that become instantaneously saturated by spin diffusion effects [21,22]. As discussed in Section 9.4.3.3, the saturation can be transferred to the labeled protein, but only at the interface, by cross-relaxation. This is observed as a reduction of peak intensity in a ^{15}N HSQC spectrum. The interface of the unlabeled protein can be identified. The interface of the second protein is easily obtained by reversing the role, with only the second protein $^{2}H/^{15}N$ uniformly labeled. As this method relies on direct through-space interactions, the identification of the interface is more precise than with CSP experiments. In particular, large conformational changes that can make it difficult to interpret CSP data do not affect cross-saturation experiments.

32.2.3 Hydrogen/Deuterium Exchange

As already addressed in Section 9.4.3.5, hydrogen/deuterium exchange is a technique that provides information on solvent accessible residues of the protein. In a deuterated solvent, amide protons on the surface of the protein exchange rapidly with deuterium, whereas the ones buried in the protein do not. If the experiment is executed on a protein complex, protons at the interface will also be protected. The hydrogen/deuterium exchange is usually followed by NMR spectroscopy using ^{15}N HSQC spectra [23] or by mass spectroscopy (MS) [24]. It reveals the solvent-accessible surface of the complex and indirectly provides the interface of the complex.

32.2.4 Intermolecular NOEs

Intermolecular NOEs provide distance restraints between pairs of atoms located at the interface. NOE restraints are very useful in protein docking. Alone, or combined with a few residual dipolar couplings (RDCs), they allow simple rigid body docking [25]. As the NOE is a through-space effect that can be measured up to 5–6 Å, intermolecular NOEs usually involve side-chain spins whose resonance assignments are typically more difficult to obtain than those for backbone atoms. In addition, the transient nature of many biomolecular complexes combined with their high molecular weight usually makes it difficult to detect such NOEs.

32.2.5 Paramagnetic Relaxation Enhancement

The paramagnetic relaxation enhancements (PRE) yields distance information between a paramagnetic center (e.g., a lanthanide ion) and the spin of interest (see Section 8.3 for a more detailed description). It can be measured on undecoupled ^{15}N HSQC spectra. This effect depends on the distance between the paramagnetic center and the nuclear spin with the same r^{-6} dependence as the NOE effect [26,27]. It accounts for the difference of line broadening between the paramagnetic and diamagnetic chemical shifts (Figure 32.1). If a paramagnetic tag is attached to one

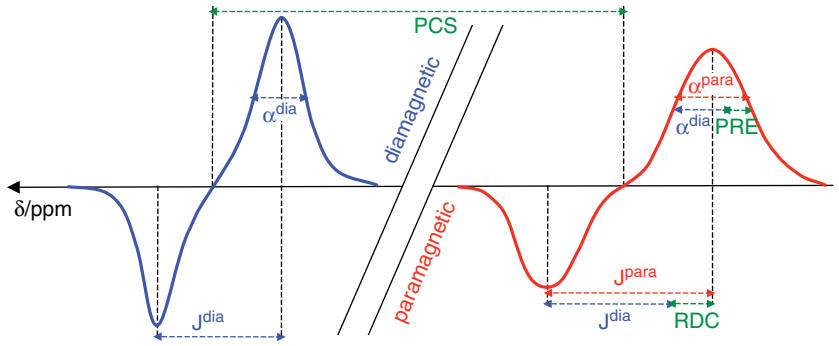


Fig. 32.1 Experimental measurement of the RDC, PRE, and PCS paramagnetic effects with two ^{15}N–^{1}H HSQC one-dimensional undecoupled spectra. The figure shows the diamagnetic and paramagnetic antiphase ^{1}H doublets. RDC is measured as the difference in line splitting. PRE can be determined from the differential line broadening. PCS is measured as the chemical shift difference.

of the protein partners, intermolecular PREs provide distance information between the paramagnetic metal ion and the second protein (see also Chapter 11), even for transient complexes [28–30]. This can be used to reduce the conformational space of the rigid-body complexes to search for. By varying the paramagnetic species (as discussed in Section 8.6; e.g., by using various lanthanides), the strength of the PRE effect can be tuned [31], allowing the measurement of very long distances up to 25–30 Å, which is particularly attractive in the context of biomolecular complexes.

32.2.6 Pseudocontact Shift

Pseudocontact shift (PCS) experiments belong to the family of paramagnetic experiments [32] requiring a paramagnetic ion attached to the protein (see also Chapter 8). PCSs are measured as the chemical shift differences between a reference (diamagnetic) spectrum and a paramagnetic spectrum recorded in the presence, for example, of a paramagnetic lanthanide ion. This concept is pictured in Figure 32.1. PCSs are usually measured in ^{15}N HSQC or ^{13}C HSQC spectra. As described in Sections 8.8.3 and 11.3.1, intramolecular PCSs allow the optimization of the $\Delta\chi$ tensor parameters of the protein to which the lanthanide is attached while intermolecular PCSs can be used to obtain the anisotropic tensor $\Delta\chi$ tensor parameters with respect to the second protein [33]. With both $\Delta\chi$ tensors being theoretically equal, they can be used to obtain the relative orientation of the two protein partners. In addition, the distance-dependence of the effect also allows the positioning of the two protein partners. The PCS effect thus provides both orientation and distance information that is in principle sufficient to perform rigid-body docking in the absence of any other energy terms or additional experimental data [34].

32.2.7 Residual Dipolar Coupling

RDCs are manifested as an increase or decrease of the magnitudes of multiplet splittings that can be observed in undecoupled NMR spectra (see also Chapter 4). This phenomenon occurs in weakly aligned systems, usually by the addition of an alignment media such as lipid bicelles, with the potential risk of altering the protein conformation [35]. As discussed in Section 8.4, an alternative way to measure RDCs is to attach a paramagnetic tag [36]; the anisotropic part of the magnetic susceptibility tensor weakly aligns the protein in the magnetic field [37]. Changes in the one-dimensional spectrum of a paramagnetic labeled protein compared to a reference (diamagnetic) spectrum are depicted in Figure 32.1. Measured and assigned RDCs can be used to obtain the alignment tensor parameters from the known structure of a protein [38,39]. In the context of complexes, this provides information to orient the two protein partners with respect to each other. This reduces the number of degree of freedom from six (three rotations + three translations) to three (three translations) in a rigid-body docking situation. Multiple alignment medias or multiple paramagnetic metal ions can be used to increase the precision of the relative orientation of the two

proteins. Recently, it has even been shown that RDCs can also provide shape information [40], allowing the complete definition of the orientation of the partners in a complex, as is the case with PCSs.

32.2.8
Diffusion Anisotropy

NMR ^{15}N relaxation rates can be used to determine the orientation of the two components of a complex with respect to each other. For ^{15}N nuclei located in secondary structure elements (i.e., rigid regions in the protein structure), the values of the relaxation rates will depend on the rotational diffusion of the protein, which can be described by a rotational diffusion tensor [41]. More specifically, the ratio between the ^{15}N transversal (R_2) and longitudinal (R_1) relaxation rates will depend on the orientation of the ^{15}N–^{1}H bond-vector with respect to the rotational diffusion tensor. As a consequence, in the case of anisotropic rotational diffusion, the R_2/R_1 rates will provide orientation information, as discussed in Section 9.4.4.1. The diffusion tensor of two components of a biomolecular complex can be determined from the experimental relaxation rates of the components within the complex. Since the rotational diffusion will be determined by the shape of the whole complex, the two components can be oriented with respect to each other in such a way that their diffusion tensors are colinear, decreasing again the number of degrees of freedom for rigid-body docking from six to three [42].

32.2.9
Non-NMR Information

Data-driven docking in HADDOCK is not limited to NMR data. Various experimental techniques can report on the overall shape of macromolecular assemblies, providing information to limit the conformational space to be explored. Small angle X-ray scattering (SAXS) experiments provide a scattering pattern that can be compared against a theoretical pattern calculated from a model of a complex [43]; for more details, see Chapter 35. Cryo-electron microscopy provides an electron density map that represents the overall shape of the complex [44]. Mass spectrometry can also provide shape information in the form of collision cross-sections measured in native ion mobility experiments [45]. Mutagenesis experiments can be performed to identify critical residues for the interaction. They consist of mutating surface residues of the protein partners and monitoring their impact on the binding – residues that do not affect binding are assumed to be located far away from the interface, while mutations that change the binding affinity are most likely located at the interface. False negatives and false positives cannot be excluded. If successful, such an investigation will report on the location of the interface of the complex, which can be used to define active and passive residues, in the same way that CSP data are used. More specific contact information can be obtained from correlated mutation experiments [46].

Finally, in the absence of any experimental information, bioinformatics interface predictions can be used (for a review, see [47]). They are typically based on a comparison of the sequence and structural features of the protein target against information contained in databases, in order to predict residues located at the interface. Several web servers are available for this purpose, some of which have been integrated in the meta predictor CPORT especially developed for use with HADDOCK [48].

32.3
How Does HADDOCK Use the Information?

Compared to other docking software packages, a unique aspect of HADDOCK is that it can handle a variety of experimental and/or predicted information to drive the

docking process and select the best models. In this section, we describe in more detail how experimental data can be translated into useful restraints for use in HADDOCK.

32.3.1 Incorporation of Ambiguous Distance Restraints

A large number of the experiments that we have presented reveal the putative interfaces in a complex without providing any specific contact or orientation information. These include CSP, mutagenesis, hydrogen/deuterium exchange, cross-saturation experiments, and interface prediction. HADDOCK was developed to exploit this ambiguous information in the form of ambiguous interaction restraints (AIRs), similar to the concept of ambiguous NOEs [49] AIRs are designed so as to bring the putative interfaces in contact during the docking, without favoring any relative orientation between the two interfaces (see Section 32.2.2). The user has to define, for each protein partner (protein A and protein B), a list of *active* and *passive* residues. The active residues are those experimentally identified to make contacts with the other protein partner, while passive residues are usually neighbor residues that might make contacts (typically not all interface residues are detected experimentally). HADDOCK uses those lists to define an *effective distance* for each active residue [5]. The effective distance of an active residue i of protein A is calculated according to:

$$d_{i\mathrm{AB}}^{\mathrm{eff}} = \left(\sum_{m_{i\mathrm{A}}=1}^{N_{\mathrm{A\ atom}}} \sum_{k=1}^{N_{\mathrm{resB}}} \sum_{n_{k\mathrm{B}=1}}^{N_{\mathrm{B\ atom}}} \frac{1}{d_{m_{i\mathrm{A}}\ n_{k\mathrm{B}}}^{6}} \right)^{\left(-\frac{1}{6}\right)} \tag{32.1}$$

where N_{Aatom} indicates all atoms of the residue i in protein A, N_{resB} indicates the active and passive residue of protein B, N_{Batom} indicates all atoms of the residue indexed by k, and d is the geometric distance between the atom $m_{i\mathrm{A}}$ and $n_{k\mathrm{B}}$. Effective distances are calculated similarly for all active residues of B. An upper limit, which is by default 2 Å, but can be user modified, is used to ensure that each protein interface faces the other: if the effective distance is larger than the defined upper limit, the active residues of each protein experience an attractive force toward the active and passive residues of the other protein. Since many atom–atom distances inversely contribute to the effective distance, an AIR restraint is typically satisfied if a residue comes within 3–5 Å of any active or passive residue of the partner molecule, depending on the number of distances entering the sum in Equation 32.1 (typically several thousands).

AIR restraints can be automatically generated from the list of active and passive residues. We have seen, however, that the determination of active residues can suffer from false negatives and false positives, such as when they are derived from chemical shift data. The use of passive residues can (partly) solve the problem of false negatives. To overcome the problem of false positives, HADDOCK automatically discards 50% (a value that can be user-modified) of the active and passive residues at random for each docking trial. Each model will thus originate from a different subset of restraints. This ensures that a large percentage of docked structures will not be calculated using wrongly defined active or passive residues. This option can be turned off if the data are of high quality and confidence.

32.3.2 Incorporation of Unambiguous Distance Restraints

NOE data provides distance information between two specific atoms (or atom groups). Although NOEs are popular for the structure determination of proteins, obtaining intermolecular NOE restraints in the case of complexes is a difficult process. Consequently, they are less often used in protein docking. However, when available, this information is highly valuable as the distance restraint is not ambiguous. HADDOCK also supports the incorporation of hydrogen bonds. The information is exploited as a harmonic distance restraint between pairs of atoms. Note that, in principle, a user is free to input restraints in any class ("ambiguous,"

"unambiguous," or "hbonds") – the main difference being that ambiguous restraints will be randomly discarded by default, while all unambiguous restraints are kept. Furthermore, the use of various classes allows the force constants for difference types of restraints to be fine-tuned.

32.3.3 Incorporation of Shape Restraints

Cryo-electron microscopy and SAXS data report on the overall shape of a complex [50]. The most common use of this kind of information is to filter out generated complexes that do not satisfy the restraints in order to improve the scoring [51,52]. They can also be used directly as restraints [53]. Due to the additional CPU requirements, they have not yet been implemented directly into HADDOCK, but at this time can only be used as a filter to increase the scoring capabilities of HADDOCK.

32.3.4 Incorporation of Orientation Restraints

RDC, PCSs, and diffusion anisotropy data are all theoretically described with a tensor: the alignment tensor for RDCs, the anisotropic tensor for PCSs, and the diffusion tensor for diffusion anisotropy data. A tensor can be seen as an orthogonal frame with different axis lengths. Each tensor frame has a specific orientation with respect to the protein assembly. RDCs are implemented in HADDOCK [54] both as intervector angle restraints [55] and as SANI restraints [56]. Diffusion anisotropy data can also be used in HADDOCK [42]. The implementation of PCS energy terms into HADDOCK [57] is based on the PARArestraints module developed by Banci *et al.* [58], which has been ported into the structure calculation software CNS. Practically, the tensor parameters can be fitted first to satisfy the experimental data, using freely available software such as PALES [39] for RDCs, Numbat [59] for PCSs, and TENSOR2 [60] or ROTDIF [61] for diffusion anisotropy data. The magnitudes of the tensor are then entered into HADDOCK as axial and rhombic components. Then, during docking, the tensor frame and protein orientations are optimized to minimize the discrepancies between measured and back-calculated data. Since HADDOCK supports side-chain and backbone flexibility, the orientation restraints can also be useful to optimize the conformation of both proteins. For this, however, it is recommended to first refine the individual components and give the refined structures as input to the docking.

32.3.5 Symmetry Restraints

If symmetry is present in the complex to be modeled (either within or between molecules), it is possible to enforce it in HADDOCK. The symmetry relationship is defined in the form of symmetry distance restraints as proposed by Nilges *et al.* [49,62]: for each restraint two distances are specified that are required to remain equal during the calculations, irrespective of the actual distance. One advantage of this method is that it is not restricted to cyclic symmetries. Other symmetries (e.g., D2 symmetry) can be enforced by various combinations of symmetry restraints. In addition, so called "noncrystallographic symmetry" restraints can defined that enforce the molecules to be identical (i.e., RMSD = 0 Å) without defining any symmetry operation between them.

32.3.6 Additional Docking Mode

HADDOCK also supports docking of proteins with DNA or RNA [63,64] and small ligands. Conformational changes in DNA can be modeled efficiently by introducing bends and twists in the nucleic acid as well as local conformational changes in the

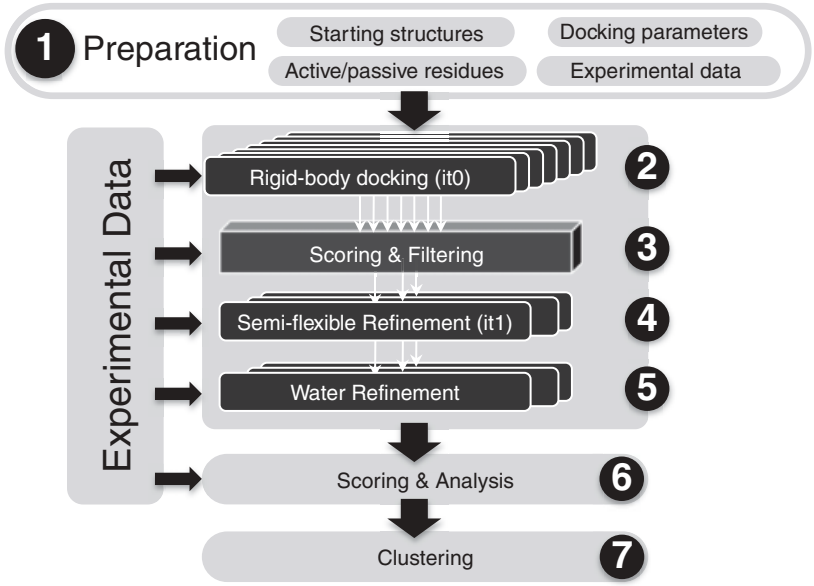


Fig. 32.2 Flow diagram of a HADDOCK run.

flexible base pairs and sugar–phosphate backbone [64]. This process is facilitated by the 3D-DART web server, which allows users to manipulate and build custom DNA conformations [65].

Multibody docking is also supported with various combinations of molecules, which now allows the assembly of up to six macromolecules together. The method has been demonstrated on a benchmark of six cases [66]. The same multibody docking approach can be used to model large domain conformational changes, as demonstrated in [67].

32.3.7 Overview of a HADDOCK Run

A typical HADDOCK run consists of seven steps (pictured in Figure 32.2):

1) The user provides the starting structure of the proteins to dock, a list of the active and passive residues for each protein, some experimental restraints, and defines some parameters for the docking.
2) HADDOCK generates 1000 rigid-body docked structures, in which the experimental data are applied to drive the docking. This is called the *it0* step. Note that the total sampling is actually 10 000 models (five models are generated per docking trial and for each the 180° symmetrical solution is also sampled, with only the best model being written to disk).
3) HADDOCK scores the models using Equation 32.2 and keeps the top 200 solutions for subsequent flexible refinement:

$$E = 0.01\,E_{\mathrm{vdW}} + 0.1\,E_{\mathrm{elec}} + 0.01\,E_{\mathrm{AIR}} - 0.01\,BSA + 1.0\,E_{\mathrm{desolv}} + 0.1\,E_{\mathrm{data}} \quad (32.2)$$

where E_{vdW}, E_{elec}, E_{AIR}, and E_{desolv} are the van der Waals, electrostatic, AIR restraint, and desolvation energies, respectively, BSA is the buried surface area, and E_{data} contains the energy of other restraint data such as NOEs, hydrogen bonds, PCSs, RDCs, diffusion anisotropy, and so on.

4) In the *it1* step, the selected models are subjected to a semiflexible refinement in torsion angle space and then scored using:

$$E = 1.0\,E_{\mathrm{vdW}} + 1.0\,E_{\mathrm{elec}} + 0.1\,E_{\mathrm{AIR}} - 0.01\,BSA + 1.0\,E_{\mathrm{desolv}} + 0.1\,E_{\mathrm{data}} \quad (32.3)$$

5) In the final *water refinement* step all models are refined in an explicit solvent shell (water (default) or dimethylsulfoxide (DMSO), depending on the type of complex – user choice) and scored using:

$$E = 1.0\,E_{\mathrm{vdW}} + 0.2\,E_{\mathrm{elec}} + 0.1\,E_{\mathrm{AIR}} + 1.0\,E_{\mathrm{desolv}} + 0.1\,E_{\mathrm{data}} \quad (32.4)$$

6) Various analysis scripts are run over the final structures (e.g., energetics analysis, hydrogen-bond and nonbonded contact analysis, restraint violation analysis, etc.)
7) The solutions are clustered using a 7.5-Å cutoff based on their pairwise ligand interface RMSD values and the cluster ranks are determined according to the average score of the four best structures of each cluster. Note that this step is performed automatically by the HADDOCK web server, while manual post-processing from the user is required if the docking is performed using a local version of HADDOCK.

32.4 Protocol: A Guided Tour of the HADDOCK Web Interface

HADDOCK is publicly available to the nonprofit scientific community either as a software package for installation or as a web interface. Installing HADDOCK on a computer or cluster requires some expertise and can only be done if other required software programs have been installed in advance. Additionally, a HADDOCK run requires many CPU hours (typically on the order of 50–100 h if run on a single CPU). To make it available to a wide community of (nonexperienced) users, HADDOCK has also been made available via a web interface[2] [4], which automatically submits the calculation to a local cluster or to worldwide Grid computing infrastructures (see WeNMR[3] and Chapter 31) [68]. The web interface has been carefully designed for ease of use for new users, while providing rich functionalities and the option of tuning specific parameters for expert users. When registering, a new user only has access to the *easy* interface, which is sufficient for standard docking. As the users progress in their understanding of and experience with HADDOCK, they can request access to the *advanced*, *expert*, and *guru* interfaces, each one of them offering increased control over HADDOCK parameters. Additional interfaces are available allowing refinement of existing complexes and multibody (up to six) docking. Access to the Grid-enabled HADDOCK portal requires separate registration and a valid personal grid certificate. In the following, we are going to explore the *guru* interface of the web server using the E2A–HPr complex as a test case; more detailed information can be found in [4] and in an online tutorial.[4]

32.4.1 Prerequisite: Registration

Before using the HADDOCK web interface, it is necessary to register for the service. To use the HADDOCK web server, go to signup page[5] and follow the instructions. The Grid-enabled version of the server distributes the calculations throughout an ensemble of interconnected computers and clusters around the world, providing access to over 10 000 CPUs to date (see the WeNMR web site). If the local HADDOCK cluster is in high use, the calculation process can become slow. We therefore strongly encourage users to use the Grid-enabled HADDOCK web server, which can be accessed from the WeNMR website under "Services." For this, a Grid certificate should be first obtained by following the instructions on the WeNMR web site (under the "access" menu) and subsequently registering with the e-NMR virtual organization (VO) as described in Section 31.3.1. This allows the use of many web services,

2) http://haddock.chem.uu.nl.
3) http://www.wenmr.eu.
4) http://haddock.chem.uu.nl/enmr/haddock-tutorial.php.
5) http://haddock.chem.uu.nl/Main/signup.html.

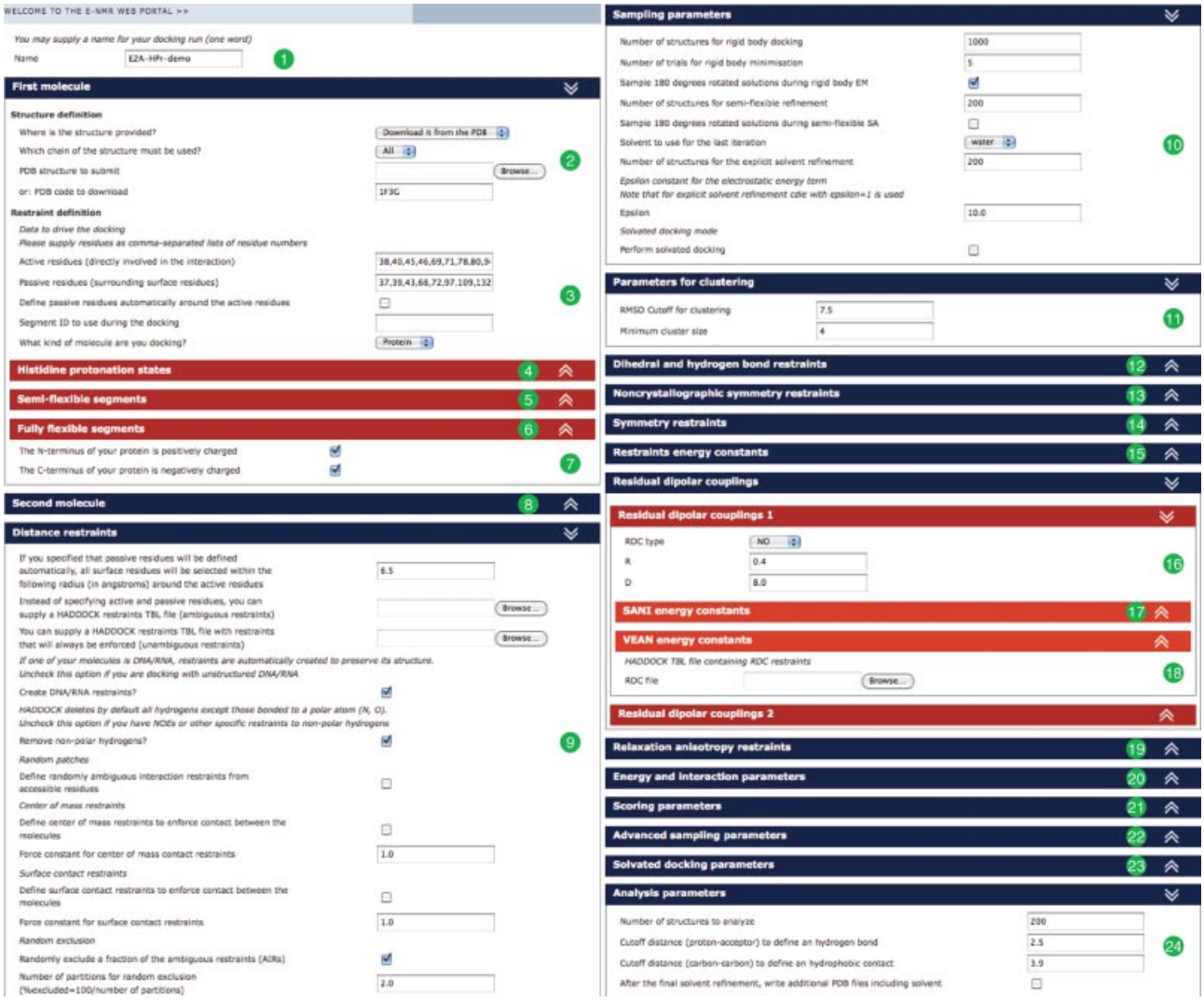


Fig. 32.3 Screenshots of the *guru* web interface for the preparation of the docking between E2A and HPr.

including HADDOCK, XPLOR-NIH [69], AMBER [70], CYANA [71], CS-ROSETTA [72], and so on, with each service requiring an independent registration.

32.4.2 Description of the Web Interface

Figure 32.3 shows two screenshots of the *guru* web interface for the preparation of the docking between the glucose-specific enzyme IIA (E2A) and the histidine-containing phosphocarrier protein (HPr). The description of the interface goes as follows:

1) Each HADDOCK session must be named by the user. In this case, we are going to dock E2A with HPr – the example name is therefore E2A-HPr-demo.
2) The user provides a PDB file containing the structure of the first protein. It can be uploaded or directly acquired from the PDB using an accession number (e.g., 1F3G for E2A). The chain ID can be provided to identify the correct protein within a large PDB file.
3) The list of active and passive residues must be entered by the user. For E2A, the following residues have been identified as active, as their resonance experiences a shift upon titration in a CSP experiment: 38, 40, 45, 46, 69, 71, 78, 80, 94, 96, and 141. These residues are on the surface of the protein, as can be verified using, for example, the program NACCESS [73]. The list of passive residues can

be manually specified, or automatically defined by HADDOCK. The type of molecule (DNA, RNA, or protein) must be specified here (when docking a small ligand, simply select "protein").

4) If known, the histidine protonation state can be manually specified. If not, by default HADDOCK will query the WHATIF server [74] to determine the most likely state (which is the case for our E2A–HPr example).
5) Up to 10 semiflexible segments can be defined by the user; however, we let HADDOCK automatically define them for our run: after the rigid-body *it0* step, intermolecular contact will be used to define semiflexible segments at the interface.
6) Up to five fully flexible segments can be defined. The difference as compared to the semiflexible segment is that fully flexible segments are made flexible throughout the entire HADDOCK protocol (except *it0*).
7) Since the termini of our protein are charged, we leave the boxes checked (default). HADDOCK will automatically update the charge for those residues. If you were to dock a subsegment from a protein, it is better to use uncharged termini (uncheck).
8) We repeat the process for the second protein: HPr. The PDB ID is 1HDN, the list of active and passive residues are 15, 16, 17, 20, 48, 49, 51, 52, 54, 56 and 11, 12, 21, 24, 40, 41, 47, 57, 85, respectively. The other parameters are identical to E2A for this run.
9) Several distance restraints can be defined. For AIRs, HADDOCK can automatically define passive residues using a cutoff distance (by default 6.5 Å). If a residue is within the cutoff and located on the surface, it will be defined as passive. For our case, we have already defined a list of passive residues, and the cutoff value has no impact.

   It is also possible to directly provide a distance restraint file (tbl format of XPLOR-NIH/CNS) for ambiguous restraints. Unambiguous restraints (e.g., for NOEs) can also be defined as a tbl file using the following syntax:

```
assign (resid 30 and segid A) (resid 12 and segid B)
4.0 1.0 2.0
```

   This will define a distance restraint of 4 Å between residue 30 of the first protein and residue 12 of the second protein. The restraint is satisfied if it is larger than $4.0 - 1.0$ Å (i.e., 3 Å) and lower than $4.0 + 2.0$ Å (i.e., 6 Å).

   The other parameters are left as default. Note the "Number of partition for random exclusion" field: a value of 2.0 will ensure that, at each docking calculation, 50% of the AIRs are discarded. This adds protection against wrongly defined active or passive residues. Statistically, many docked structures will be calculated without any false-negative or false-positive interface residue definition.
10) The sampling parameters section controls the number of structures calculated. The default value is 1000 structures generated at *it0*, of which only 200 are subject to the *it1* and water refinement stages.
11) The RMSD cutoff for the clustering is set by default to 7.5 Å. A smaller value will result in more clusters and should be used if all generated structures have a small interface RMSD with respect to each other. HADDOCK disregards clusters with less than four structures.
12) Dihedral and hydrogen bonds restraints can be provided in the form of a tbl file.
13) and 14) Allow the definition of symmetry restraints for the system (both between and within molecules).
15) The "restraints energy constant" section allows the modification of the relative weight of the ambiguous, unambiguous, and hydrogen bond distance restraints. The default values are optimized for most cases.
16) RDCs can be specified in two different forms: SANI [56] or as intervector angle restraints [55] (VEAN). In both cases, one must specify the axial (D) and rhombic (R) components of the alignment tensor.

17) For SANI restraints, one must provide a restraint file defining the restraints with respect to an alignment tensor. For each RDC, the syntax is as follows:

```
assign (resid 995 and resname ANI and name OO)
(resid 995 and resname ANI and name Z)
(resid 995 and resname ANI and name X)
(resid 995 and resname ANI and name Y)
(segid B and resid 2 and name N)
(segid B and resid 2 and name HN) 12.1 0.2
```

The first four lines identify the tensor, and only the residue number should be updated to correspond to one of the five datasets allowed: 995 for the first dataset, 996 for the second, . . ., 999 for the last dataset. The last two lines correspond to the RDC of residue 2, with the actual RDC value set to 12.1 Hz and the tolerance to 0.2 Hz.

18) In the case of VEAN restraints, the user must upload an intervector projection angle restraints file. The generation of these restraints is implemented in a slightly modified version of a Python script (dipolar_segid.py) kindly provided by Dr. Helen Mott and Dr. Wayne Boucher (Cambridge University). This script is distributed with the HADDOCK program. The PCS energy term is not yet implemented in the web server version of HADDOCK and requires installation of the software version [57].
19) If available, anisotropy restraints (DANI) can be provided. This comprises the axial (D) and rhombic (R) components of the diffusion tensor, a restraint tbl file (syntax similar to RDC-SANI), and the proton/nitrogen frequencies.
20) The "Energy and interaction parameters" section controls various energy parameters used during calculations. We will keep the default values for our example calculation.
21) In the "Scoring parameters" section, the user can control the relative weight of all energy terms for each HADDOCK stage. While the default weight for the van der Waals energy, electrostatic energy, desolvation energy and *BSA* score are optimized for general use, the user might want to change the weight for data-related energy terms to reflect the confidence in the quality and value of the data.
22) The advanced sampling parameters are of interest only for advanced users. These allow control of the number of steps of the minimization algorithm and control of the temperatures during simulated annealing. The default values are fine for most cases.
23) Solvated docking provides a way to include water molecules in all stages of the docking process. It must be activated in the sampling parameters section (10). Solvated docking is much slower than standard docking. More details on the role of the parameters can be found in [75].
24) At the end of the docking, HADDOCK automatically performs some analysis, including a hydrogen bond and hydrophobic contacts count. The cutoff distance used for counting can be specified here.

32.4.3 Analysis of the Docking Run

The calculations usually take between a few hours (typically on the web server) and a few days, depending on the size of the system and number of processors available. Upon completion, HADDOCK automatically sends an e-mail with a link to retrieve the results. Information about the different clusters is conveniently displayed on a web page. This includes the HADDOCK final score, the size of the cluster of structures, and the value of the different energy terms. It is possible to directly visualize the top structures of each cluster, to download the individual structures for further analysis, or even to download the complete run as a gzipped tar archive.

32.5 Troubleshooting

This section provides solutions for some common problems encountered by users.

32.5.1 General Considerations

Users might be interested in using HADDOCK for several different types of investigations:

i) Using *homology models* with HADDOCK can be performed, but the results strongly depend on the quality of the model.
ii) *Up to six proteins* can be docked together; for more details see [66].
iii) Docked structures (from other programs) can be refined with HADDOCK. This can be done in the *refinement web interface*.[6] For this, the components of the complex should be uploaded separately. Their relative orientation will be maintained.
iv) HADDOCK can run even if *only one protein has experimental data*. In this case, all the surface residues for the second partner must be defined as passive. Never define all residues as passive since this will result in an excessive number of distances to be calculated for the ambiguous distance restraints.
v) HADDOCK can run even in *ab initio* mode, when no data are provided. The success of the docking strongly depends on the quality of the starting structures, shape, and possible symmetry considerations. In this case, center of mass restraints should be activated (in the *expert interface*). Alternatively (and recommended), interface predictions could be used to define active residues.

32.5.2 Problems Related to the PDB File

For a PDB file to be accepted by HADDOCK, various requirements must be met:

i) All PDBs should end with an END statement and have no SEGID (columns 73–76).
ii) For *NMR* ensembles, all models must be identical (except for the coordinate numbers).
iii) Each protein must not have *overlapping residues*. In particular, when dealing with multichain proteins, the residue number must be different.
iv) Crystal structures containing *multiple occupancy* for atoms/residues are not supported by HADDOCK. It should be converted in an ensemble of structures, or only one occupancy state should be kept
v) *Elemental ions* can be included in HADDOCK, but their proper charge must be specified in both the residue name and the atom name. For example, for Fe^{3+}, the atom name and residue name must be "FE+3" and "FE3," respectively.
vi) *Cofactors* must be specified in the PDB as HETATM, and it is recommended to add a TER statement between the chains and submolecules.
vii) *Modified amino acids* must be specified in the PDB as ATOM. Only modified amino acids listed at the HADDOCK library[7] are supported.
viii) Starting structures for double-stranded *DNA* can be easily generated using the 3D-DART web server developed by our group [65].

[6] http://haddock.chem.uu.nl/services/HADDOCK/haddockserver-refinement.html.
[7] http://haddock.chem.uu.nl/services/HADDOCK/library.html.

32.5.3 Problems Encountered During Docking

Errors can occur during the different stages of HADDOCK calculation. In most cases, the error message will provide some guidance on how to solve the problem.

i) During the *topology generation step* or the *rigid-body stage*, errors are most likely caused by low-quality starting structures or by a problem with cofactors.
ii) Errors during *the simulated annealing* or *water refinement stage* are usually due to poor-quality starting structures that have "exploded," or by cofactors, ions, or protein fragments that have drifted away.
iii) Errors during the *clustering of solutions* indicates that the generated structures are too dissimilar. While using a larger cutoff for clustering is a short-term fix, the real solution is to try to improve the quality of the starting structures and/or to incorporate more experimental restraints.

Further Reading

Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003) HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731–1737.

Bonvin, A.M.J.J. (2006) Flexible protein–protein docking. *Curr. Opin. Struct. Biol.*, **16**, 194–200.

van Dijk, M., van Dijk, A.D., Hsu, V., Boelens, R., and Bonvin, A.M.J.J. (2006) Information-driven protein–DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res.*, **34**, 3317–3325.

de Vries, S.J., van Dijk, A.D.J., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A.M.J.J. (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins*, **69**, 726–733.

de Vries, S.J. and Bonvin, A.M.J.J. (2008) How proteins get in touch: interface prediction in the study of biomolecular complexes. *Curr. Protein Pept. Sci.*, **9**, 394–406.

Zacharias, M. (2008) Combining elastic network analysis and molecular dynamics simulations by Hamiltonian replica exchange. *J. Chem. Theory Comput.*, **4**, 477–487.

Vajda, S. and Kozakov, D. (2009) Convergence and combination of methods in protein–protein docking. *Curr. Opin. Struct. Biol.*, **19**, 164–170.

de Vries, S.J., van Dijk, M., and Bonvin, A.M.J.J. (2010) The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.*, **5**, 883–897.

Janin, J. (2010) Protein–protein docking tested in blind predictions: the CAPRI experiment. *Mol. Biosyst.*, **6**, 2351–2362.

Karaca, E., Melquiond, A.S.J., de Vries, S.J., Kastritis, P.L., and Bonvin, A.M.J.J. (2010) Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server. *Mol. Cell Proteomics*, **9**, 1784–1794.

Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2010) Protein–protein docking dealing with the unknown. *J. Comput. Chem.*, **31**, 317–342.

van Dijk, M. and Bonvin, A.M.J.J. (2010) Pushing the limits of what is achievable in protein–DNA docking: benchmarking HADDOCK's performance. *Nucleic Acids Res.*, **38**, 5634–5647.

Karaca, E. and Bonvin, A.M.J.J. (2011) A multi-domain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. *Structure*, **19**, 555–565.