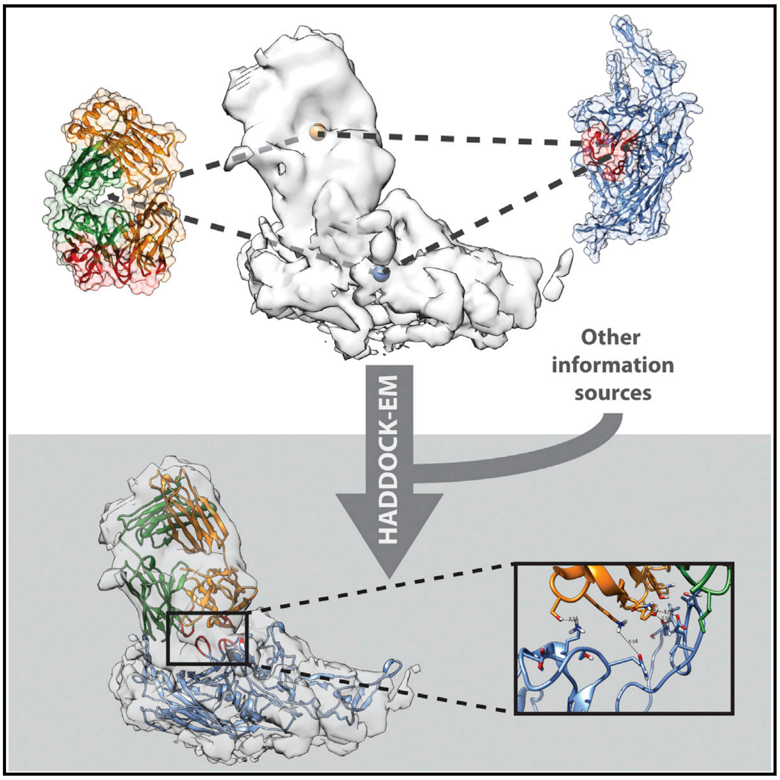
Structure

Resource

Integrative Modeling of Biomolecular Complexes: HADDOCKing with Cryo-Electron Microscopy Data

Graphical Abstract



Highlights

- HADDOCK now supports the use of cryo-EM data to drive the docking and score models
- Cryo-EM data can be combined with all other sources of data available in HADDOCK
- Cryo-EM restraints drive conformational changes during flexible refinement
- The resulting models have chemically relevant interfaces revealing new interactions

Authors

Gydo C.P. van Zundert, Adrien S.J. Melquiond, Alexandre M.J.J. Bonvin

Correspondence

a.m.j.j.bonvin@uu.nl

In Brief

Van Zundert et al. incorporate cryo-EM data into the data-driven flexible docking program HADDOCK. The approach is validated and illustrated on five experimental cryo-EM datasets, including examples of integrative modeling with additional interface information (mutagenesis and hydroxyl radical footprinting data). The resulting models have high-quality interfaces, revealing novel details of the interactions.

van Zundert et al., 2015, Structure *23*, 949–960
May 5, 2015 ©2015 Elsevier Ltd All rights reserved
http://dx.doi.org/10.1016/j.str.2015.03.014

Integrative Modeling of Biomolecular Complexes: HADDOCKing with Cryo-Electron Microscopy Data

Gydo C.P. van Zundert,[1] Adrien S.J. Melquiond,[1,2] and Alexandre M.J.J. Bonvin[1,*]
[1]Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht 3584CH, the Netherlands
[2]Present address: Hubrecht Institute - KNAW & University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, the Netherlands
*Correspondence: a.m.j.j.bonvin@uu.nl

http://dx.doi.org/10.1016/j.str.2015.03.014

SUMMARY

Protein-protein interactions play a central role in all cellular processes. Insight into their atomic architecture is therefore of paramount importance. Cryo-electron microscopy (cryo-EM) is capable of directly imaging large macromolecular complexes. Unfortunately, the resolution is usually not sufficient for a direct atomic interpretation. To overcome this, cryo-EM data are often combined with high-resolution atomic structures. However, current computational approaches typically do not include information from other experimental sources nor a proper physico-chemical description of the interfaces. Here we describe the integration of cryo-EM data into our data-driven docking program HADDOCK and its performance on a benchmark of 17 complexes. The approach is demonstrated on five systems using experimental cryo-EM data in the range of 8.5–21 Å resolution. For several cases, cryo-EM data are integrated with additional interface information, e.g. mutagenesis and hydroxyl radical footprinting data. The resulting models have high-quality interfaces, revealing novel details of the interactions.

INTRODUCTION

Protein interactions underlie most of the complexities encountered in the cell. They play a determining role in processes ranging from protein translation to muscle contraction. Numerous diseases are the result of mutations at the interface of protein complexes (Joerger and Fersht, 2007; Lage, 2014). For a thorough and fundamental understanding of these processes and rational drug design, knowledge of these interactions and interfaces at an atomic level is of paramount importance (Wells and McClendon, 2007; Nero et al., 2014). Unfortunately, the number of available high-resolution structures of protein complexes determined by either X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy remains rather sparse compared with the size of the interactome (Mosca et al., 2013; Petrey and Honig, 2014).

Cryo-electron microscopy (cryo-EM) is a technique capable of imaging large biomolecular complexes in their native hydrated state (Orlova and Saibil, 2011). The resolution is, however, usually limited to such an extent that a direct atomic view of the interface is out of the question. In order to remedy this, cryo-EM data are often combined with high-resolution atomic structures (Esquivel-Rodríguez and Kihara, 2013). The simplest and most common way of building macromolecular assemblies into cryo-EM maps is by manual fitting of atomic structures using dedicated graphics software (Baker and Johnson, 1996; Goddard et al., 2007). A more objective but less used method is full-exhaustive search rigid-body fitting, for which a plethora of software has been developed as reviewed in Esquivel-Rodríguez and Kihara (2013). Still, as the resolution decreases, placement of subunits becomes ambiguous and more models need to be sampled and/or additional data incorporated into the modeling to generate sensible models.

Protein-protein docking is in principle well suited for this task (Moreira et al., 2010; Huang, 2014), since it naturally samples a large number of conformations and can take into account additional sources of information for scoring and/or for driving the docking process (Karaca and Bonvin, 2013a; Rodrigues and Bonvin, 2014). Several docking programs have incorporated cryo-EM data into their work flow. MultiFit automatically segments the cryo-EM density using a Gaussian mixture model to deduce anchors, subsequently docking the components of the complex onto the anchors (Lasker et al., 2010). EMLZerD uses the cryo-EM data to score the models using 3D Zernike descriptors (Esquivel-Rodríguez and Kihara, 2012). A recent approach has been implemented in ATTRACT-EM (De Vries and Zacharias, 2012), which represents the cryo-EM data by a Gaussian mixture model and fits the subunits into the map in a procedure reminiscent of Kawabata's approach (Kawabata, 2008); the resulting models are then refined. Most of these methods, however, separate the use of the cryo-EM data from the use of other sources of information: They first fit the structures in the density and only afterward might take into account the physico-chemical properties (energetics) of the interface. Furthermore, they usually do not actively use additional orthogonal information that may be available, such as mutagenesis or mass spectrometry cross-link data.

Only a few approaches have been published that can incorporate a variety of data (Alber et al., 2008), one of which is the Integrative Modeling Platform (IMP) developed by the Sali group, which has the capability of integrating cryo-EM data among others (Topf et al., 2008; Schneidman-Duhovny et al., 2012; Velázquez-Muriel et al., 2012). Another approach is our in-house data-driven docking software HADDOCK (Dominguez et al., 2003; De Vries et al., 2010a), which is already capable of actively using information from various sources, such as mutagenesis, NMR H/D exchange and cross-links data, to name only a few. In addition, it is able to deal with multiple subunits (Karaca

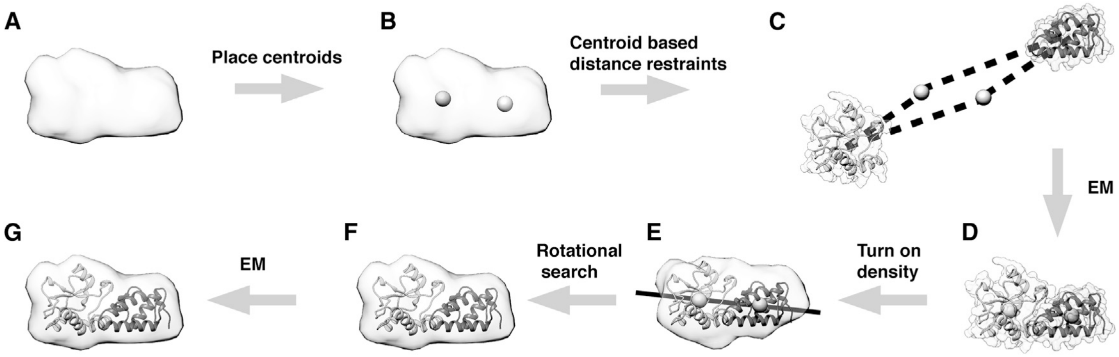


Figure 1. Representation of the Rigid-Body Docking Protocol in HADDOCK-EM

(A) Simulated 15-Å cryo-EM data of the 7CEI complex.

(B) The density with centroids (gray spheres) representing the approximate center of mass (COM) of each subunit.

(C) Initial docking setup in HADDOCK. Distance restraints are defined between the COM of chain A (light gray) and B (dark gray) of 7CEI and their corresponding centroids.

(D) An initial complex is formed after a rigid-body energy minimization (EM).

(E) The position of the subunits is approximately correct but their orientation in the cryo-EM map should still be determined.

(F) A fine rotational search is performed around the axis that is formed by the line joining the two centroids. The orientation with the highest cross-correlation value is chosen.

(G) A final rigid-body EM is performed now directly against the cryo-EM data using a cross-correlation-based potential without the centroid-based distance restraints.

et al., 2010), can handle proteins, peptides (Trellet et al., 2013), DNA (Van Dijk et al., 2006), and RNA complexes and any combination thereof. HADDOCK leverages its unique ability to combine multiple structural data into the modeling process to implement powerful strategies to deal with large domain conformational changes (Karaca and Bonvin, 2011). Here we describe how we have incorporated cryo-EM data into HADDOCK, such that the density is actively used as an additional energy term during docking, scoring, and flexible refinement. These cryo-EM restraints can be combined with all other already available sources of information and restraints supported in HADDOCK. We first report on the optimization and benchmarking of our method on 17 complexes from the protein-protein docking benchmark 4.0 (Hwang et al., 2010) using simulated data of 10, 15 and 20 Å, and a multi-component symmetrical complex. Then we demonstrate its applicability in five cases using available experimental data for two ribosome complexes, based on 9.8 Å (Guo et al., 2011) and 13.5 Å (Boehringer et al., 2012) data; two virus-antibody complexes using 8.5 Å (McCraw et al., 2012) and 21 Å (Wang et al., 2013) resolution data; and a symmetric pentamer using 16 Å negative stain data (Daudén et al., 2013). In several cases, additional interface information is included based on mutagenesis data and the biology of the system. The resulting models have high-quality interfaces without the clashes usually found in manually fitted models, revealing new details of the interactions.

RESULTS AND DISCUSSION

Implementation of Cryo-EM Data into HADDOCK

We first describe the implementation of cryo-EM restraints into the rigid-body docking stage of HADDOCK (HADDOCK-EM, Figure 1). The approximate position of the center of mass (COM) of each chain in the density map is represented by a centroid. The positions of these centroids can be determined in multiple ways: subunits can first be placed manually in the density in the correct position after which the COM can be calculated; a full-exhaustive cross-correlation search of the chains in the density using rigid-body fitting software can be used (e.g. Hoang et al., 2013) to extract positions corresponding to high cross-correlation values; several automatic methods have been devised for simultaneous centroid placement (Birmanns and Wriggers, 2007; Lasker et al., 2010; Wriggers et al., 1998; Zhang et al., 2010); a more elaborate approach combines cross-link data with the cryo-EM map to infer the positions of the subunits (Murakami et al., 2013).

Once the centroids have been determined, the docking can start. First the chains are separated in space at an approximate minimal distance of 25 Å from each other and given a random orientation. Distance restraints are defined between the COM of each protein to either a specific centroid if one is able to distinguish the two chains in the density, or ambiguously to all centroids if the chains cannot be distinguished in the density. The former can be interpreted as unambiguous and the latter as ambiguous distance restraints. We thus transform the density data into distance restraints for several reasons. First and foremost, this increases the radius of convergence of pulling the chains into the density toward specified positions compared with using a cross-correlation potential, making the approach more robust. Indeed, when using only the cross-correlation, we found that the chains often get stuck in local minima before they can even interact with each other. Second, the distance restraints approach falls within the original philosophy of HADDOCK, making it easier to combine cryo-EM data with other relevant information sources. Having defined the cryo-EM-derived distance restraints, we then dock the initial complex by means of rigid-body energy minimization, which effectively positions it into the cryo-EM map to fit the centroids. In the case of binary complexes, the optimal orientation of the complex with

Table 1. Description of the Complexes in the Benchmark

PDB Code	Category[a]	Difficulty[b]	i-RMSD[c]	Residues A	Residues B
1AVX	E	Easy	0.47	233	176
2OUL	E	Easy	0.53	241	110
1AY7	E	Easy	0.54	96	89
4CPA	E	Easy	0.62	307	39
1AHW	A	Easy	0.69	428	206
7CEI	E	Easy	0.70	130	87
2OOB	O	Easy	0.85	41	71
2FD6	A	Easy	1.07	428	279
1AK4	O	Easy	1.33	164	137
1B6C	O	Easy	1.96	329	107
1BGX	A	Medium	1.48	822	423
1R6Q	O	Medium	1.67	141	89
1M10	E	Medium	2.10	266	207
1ACB	E	Medium	2.26	245	70
1JK9	O	Hard	2.51	220	153
1BKD	O	Hard	2.86	479	166
1JMO	O	Hard	3.21	385	280

The 17 protein-protein complexes used during the optimization and benchmarking of HADDOCK-EM. The complexes were taken from the protein-protein docking benchmark 4.0 (Hwang et al., 2010).

[a] The category of the complex: E, enzyme/inhibitor or enzyme/substrate; A, antibody/antigen; O, others.

[b] The difficulty of the complex according to the CAPRI standard.

[c] i-RMSD: RMSD of Cα atoms of interface residues calculated after finding the best superposition of bound and unbound interfaces.

respect to the density still needs to be determined since the centroid-based docking allows for rotational ambiguity. Therefore, we perform a fine rotational search of the complex around the axis formed by the line joining the centroids and score each orientation using the cross-correlation value between the model and the map. The orientation corresponding to the highest cross-correlation value is further refined using rigid-body energy minimization where the energy consists of the nonbonded interaction terms of classical force fields (intermolecular van der Waals and electrostatic energies) and an added cross-correlation potential. Typically 10,000 solutions are generated at the rigid-body docking stage. All calculations are performed with CNS (Crystallography and NMR System) (Brunger, 2007) (see the Experimental Procedures section for details).

After the rigid-body stage, the generated solutions are scored with the HADDOCK-EM-it0 score, which corresponds to the original HADDOCK score (see Equation 1) complemented with a local cross-correlation-based energy (see the Experimental Procedures; Equations 6 and 7). The 400 best-scoring models are then refined using the standard HADDOCK refinement protocol with an additional correlation-based potential to further fit the chains into the density, while reckoning with the energetics of the system.

Impact of Cryo-EM Data in the Rigid-Body Docking Stage

Since the HADDOCK protocol consists of several stages (rigid-body docking and scoring [it0] and flexible refinement stages in a vacuum [it1] and explicit solvent [itw]), we discuss the impact of incorporation of cryo-EM data on each stage separately. We investigated the use of 10-, 15-, and 20-Å simulated cryo-EM data on a benchmark consisting of 17 complexes taken from the protein-protein docking benchmark 4.0 (Hwang et al., 2010). These complexes consist of ten easy, four medium, and three hard cases (based on the degree of conformational changes taking place upon complex formation) and are listed in Table 1. Even though the complexes in the benchmark are significantly smaller than what can be imaged by cryo-EM, their use is still justified to optimize our protocol and investigate the limits of using density data during the docking.

As a reference to assess the performance of using cryo-EM data in the it0 stage, we used the ab initio mode of HADDOCK (HADDOCK-CM), which uses COM distance restraints between molecules to drive the docking (Karaca and Bonvin, 2013b). We investigated two different performance indices at this stage, namely the interfacial quality of the best-generated solution, or interface root-mean-square deviation (i-RMSD) as defined by the CAPRI standards (Janin et al., 2003), and the number of acceptable solutions at the rigid-body docking stage among the 10,000 models generated. We define an acceptable solution as having an i-RMSD $\leq$ 4.0 Å from the native complex.

As can be seen in Figure 2A, HADDOCK-CM generates at the rigid-body stage at least one acceptable solution of 10,000 in 11 of the 17 cases, of which nine are from easy and two are from medium difficulty targets. The HADDOCK-EM protocol generates at least one acceptable solution in 13 of 17 cases, independent of the resolution of the simulated density maps used for the docking, of which ten were easy targets, two were medium and one was a hard target. The quality of the best-generated model improves for all complexes compared with HADDOCK-CM, except for the smallest 2OOB complex when using 15- and 20-Å resolution data. The average i-RMSD improvement is 1.2, 1.5, and 1.6 Å when using 20-, 15-, and 10-Å data, respectively. Even for complexes for which no acceptable solutions were generated, there is a considerable increase of quality, e.g. the i-RMSD of the hard 1JMO complex decreases from 6.13 Å for HADDOCK-CM to 4.66, 4.35, and 4.51 Å when using 20-, 15-, and 10-Å data, respectively.

Moreover, not only the quality of the interface of the best model benefits from the use of cryo-EM data, but the number of acceptable solutions generated also increases significantly (Figure 2B). For HADDOCK-CM, the median number of acceptable solutions generated is 1, while for HADDOCK-EM it increases to 8, 17, and 46 when using 20-, 15-, and 10-Å data, respectively. The only complex where HADDOCK-CM actually generates more acceptable solutions compared with HADDOCK-EM is again the small globular 2OOB complex.

As our protocol is dependent on the input of centroid coordinates, we also investigated its sensitivity to incorrect centroid placement. To this end, we repeated the docking for five cases where both centroids were separately displaced by 3, 5, and 7 Å in a random direction. The total error in placement was thus 14 Å total in the latter case. The difference in the number of acceptable solutions generated in the top 400 differed per case (see Table S2). Only at 7-Å displacement of both centroids does the number of acceptable solutions in the top 400 decrease consistently, but is still significantly larger compared with

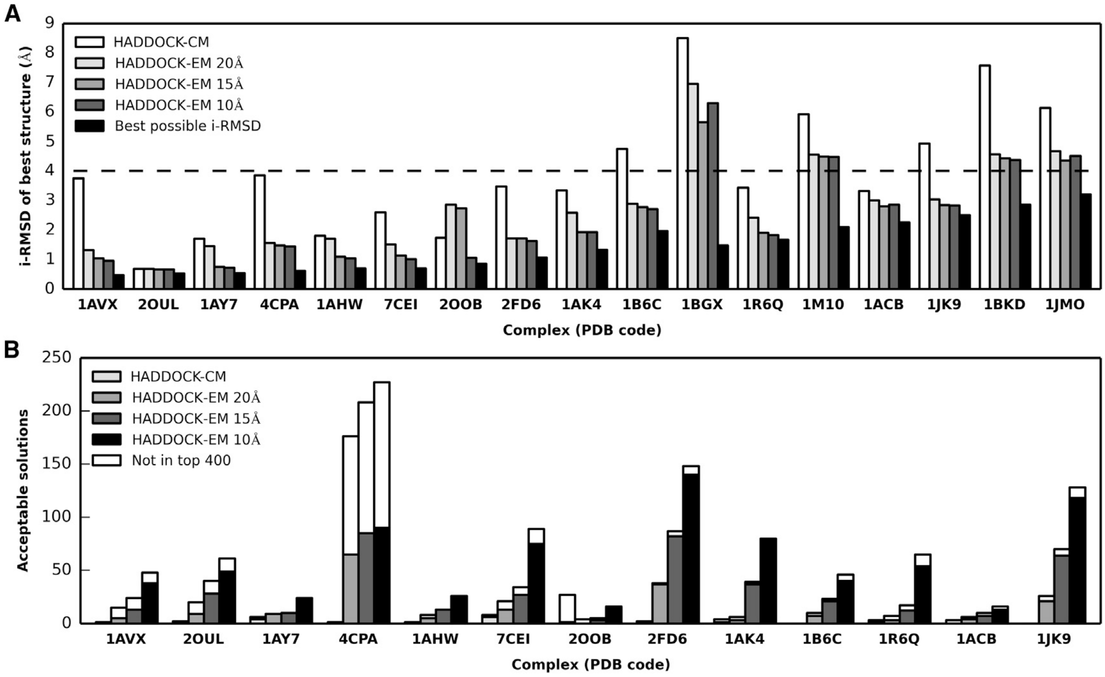


Figure 2. Quality and Number of Acceptable Models Generated after Rigid-Body Docking

(A) Interface RMSD (i-RMSD) of the best model generated after the it0 stage for the 17 complexes of the benchmark. White bar, HADDOCK-CM (ab initio docking mode with center of mass restraints); light gray, gray, and dark gray bar, HADDOCK-EM using 20-, 15-, and 10-Å simulated cryo-EM data, respectively; black bar, minimal i-RMSD of unbound compared with bound complex. The complexes are ordered according to their difficulty level. The dashed line represents the cutoff for a solution to be acceptable (i-RMSD $\leq$4 Å).

(B) Number of acceptable solutions generated after the it0 stage and the number of acceptable solutions in the 400 best-scoring models. The height of each bar represents the number of acceptable solutions generated in the 10,000 models; the height of the inner solid bars represents the number of acceptable solutions that are in the top 400 after scoring. Only complexes for which acceptable solutions were generated are displayed. Light gray, HADDOCK-CM; gray, dark gray, and black bar, HADDOCK-EM using 20-, 15-, and 10-Å simulated cryo-EM data.

HADDOCK-CM. Thus, our approach is robust against centroid placement errors up to at least 7 Å.

Impact of Cryo-EM Data on the Scoring of Rigid-Body Docking Solutions

To incorporate the cryo-EM data into the scoring function, we supplemented the original HADDOCK score with a local cross-correlation (LCC) energy term (HADDOCK-EM score). The efficiency of this combined score is shown in Figure 2B. The HADDOCK-CM models were scored with the original HADDOCK score, which resulted in at least one acceptable solution in the top 400 models for 7 of the 11 successful cases, where at least one acceptable solution was generated. The HADDOCK-EM models were scored with the HADDOCK-EM score, which resulted in at least one acceptable solution for all 13 successful cases, irrespective of resolution, with the exception of the 2OOB complex using 20-Å resolution data.

To investigate the effect of the LCC term in the HADDOCK score, the total number of acceptable solutions in the top 400 was calculated for the HADDOCK-EM models using the regular HADDOCK and HADDOCK-EM score. The influence of the LCC term in the HADDOCK score is significant, as the median number of acceptable solutions in the top 400 increases from 3 to 5, 4 to 13, and 13 to 38 when using 20-, 15-, and 10-Å data, respectively. The HADDOCK-EM score is able to rank 52%, 69%, and 78% of the generated acceptable solutions in the top 400 compared with 38%, 39%, and 41% when using the regular HADDOCK score with 20-, 15-, and 10-Å resolution data, respectively.

The discriminative ability of the LCC term increases with the resolution, as expected. When plotting the LCC versus the i-RMSD (see Figure S1), we observe a funnel shape for most complexes, with high LCC values found for complexes with low i-RMSD values. This becomes even more pronounced as the resolution of the data increases. For higher i-RMSD, the correlation is lost and the LCC is no longer indicative of the quality of the solutions as was observed before (Shacham et al., 2007). It should further be noted that the absolute value of the LCC term is not indicative of the quality of the model. For example, when using 20-Å data, correlation values of >0.9 are routinely found for non-native models. As such, the correlation value only has meaning in a comparative setting, highlighting the need to sample and score multiple conformations.

Effect of Cryo-EM Data on the Flexible Refinement Stage

Next we investigated the impact of incorporating cryo-EM restraints on the flexible refinement stage of HADDOCK. We calculated the i-RMSD improvement of the 400 best-scoring it0 models after each refinement stage for all complexes. A

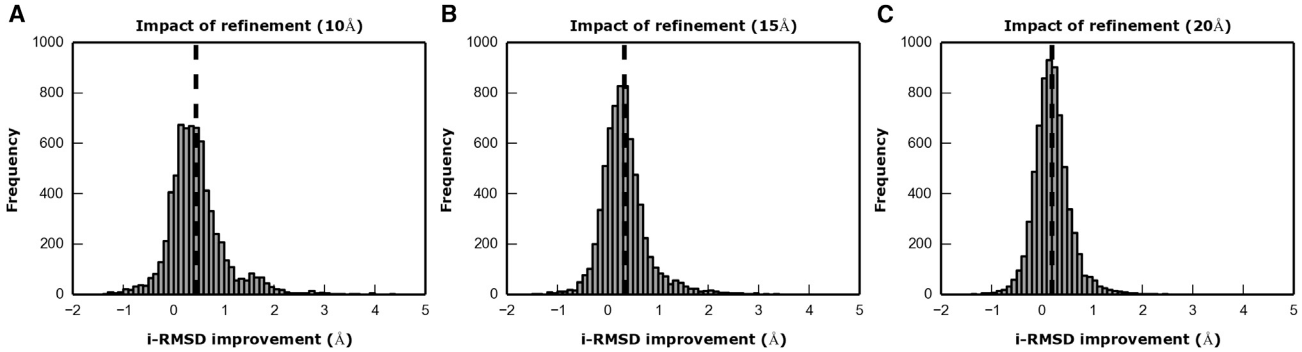


Figure 3. Effect of the Flexible Refinement Stage with Cryo-EM Restraints on i-RMSD

The i-RMSD improvement (i-RMSD it0–i-RMSD itw) for all refined complexes after itw when using 10- (A), 15- (B), and 20-Å (C) data plotted as a histogram. Positive values indicate a decrease in i-RMSD toward the native structure. The dashed vertical line in the figures represents the average i-RMSD improvement.

histogram of i-RMSD improvements after the it0 and itw stages is shown in Figure 3. The average i-RMSD improvement after refinement when using 20-Å data is 0.20 Å with a maximum of 2.49 Å. This increases to an average of 0.33 and 0.45 Å and a maximum of 3.34 and 4.37 Å when using 15- and 10-Å data, respectively. The average i-RMSD improvements between it1 and itw are modest: 0.04, 0.05, and 0.10 Å when using 20-, 15-, and 10-Å data with maximums of 0.25, 0.34, and 0.43 (see Figure S2). So the bulk of the improvement is gained during the it1 stage, which was also previously noted (see Figure 2 in De Vries et al., 2007). The maximal improvement observed with cryo-EM restraints is about two times larger than what was previously observed in an analysis of our CAPRI predictions. This substantial improvement is also reflected in the increased number of acceptable solutions after the refinement for each complex (Table S1). The number of cases with at least one acceptable solution increases from 13 for 20-Å resolution data to 15 for the 15- and 10-Å resolution data (Table 2). The resulting models are ultimately re-scored using the itw-HADDOCK-EM score (Equation 7). The enrichment of models in the top 400, 10, and 1 compared with HADDOCK-CM are given in Table S3.

Table 2. i-RMSD Values of the Best-Generated Complex after the Rigid-Body Docking and Final Water Refinement Stages

	i-RMSD of Best Solution (Å)					
	20-Å Data		15-Å Data		10-Å Data	
PDB Code	It0	Itw	It0	Itw	It0	Itw
1AVX	1.32	1.20	1.04	0.84	0.96	0.67
2OUL	0.68	0.89	0.66	0.70	0.66	0.63
1AY7	1.45	1.05	0.75	0.73	0.72	0.66
4CPA	1.55	1.53	1.48	1.24	1.44	0.94
1AHW	1.70	1.05	1.09	1.03	1.04	0.91
7CEI	1.51	1.50	1.14	0.91	1.01	0.78
2OOB	2.85	4.04	2.73	2.31	1.06	0.97
2FD6	1.71	1.36	1.71	1.17	1.62	1.13
1AK4	2.58	2.58	1.93	1.54	1.93	1.22
1B6C	2.89	2.33	2.78	2.09	2.71	1.88
1BGX	6.95	5.42	5.65	4.07	6.29	4.85
1R6Q	2.41	2.74	1.91	1.68	1.83	1.26
1M10	4.55	3.81	4.48	3.20	4.47	2.82
1ACB	3.00	2.73	2.80	2.45	2.86	2.43
1JK9	3.04	2.83	2.84	2.37	2.83	2.32
1BKD	4.56	4.21	4.43	3.82	4.37	3.62
1JMO	4.66	4.55	4.35	4.20	4.51	4.23

The quality in terms of i-RMSD[a] values of the best solution generated after it0 and itw stages is given for each complex at the three cryo-EM density resolutions.

[a] The i-RMSD is calculated by fitting the solution on the backbone atoms of the residues involved in intermolecular contacts in the native complex within a cutoff of 10 Å.

Docking Two Ribosomal Proteins Using Experimental 9.8-Å Cryo-EM Data

As a test case using experimental cryo-EM data, we docked the S7 and S19 proteins of the 30S *E. coli* ribosome using a 9.8-Å cryo-EM map (EMD-1884). The map has a corresponding atomic structure (2ykr), which has been modeled by manually fitting a crystal structure of the full ribosome in the map as a rigid body, shown in Figure 4A.

We docked the two proteins using only the fraction of the cryo-EM density that can be attributed to the two proteins. The centroids were determined by calculating the position of the COM of each protein as they were currently placed in the density (Figure 4B). Applying HADDOCK-EM resulted in 15 clusters, with the best-scoring cluster containing 105 of the 400 generated solutions of which the best-scoring complex has an i-RMSD of 1.56 Å compared with the crystal structure (Figure 4C).

Integrative Modeling of KsgA with rRNA Using 13.5-Å Cryo-EM Data

As a more realistic example, we applied HADDOCK-EM to model the binding of KsgA, a methyltransferase, to the 30S maturing *E. coli* ribosome. Crystal structures are available for the 30S ribosome and KsgA together with a 13.5-Å cryo-EM map of the complex (EMD-2017). The rRNA can be unambiguously fitted in the density because of the higher density of the phosphates in the backbone. The cryo-EM data clearly show the density of KsgA, revealing that helices 24, 27, and 45 of the rRNA are involved in the interaction (Figure 5A), which has been corroborated by

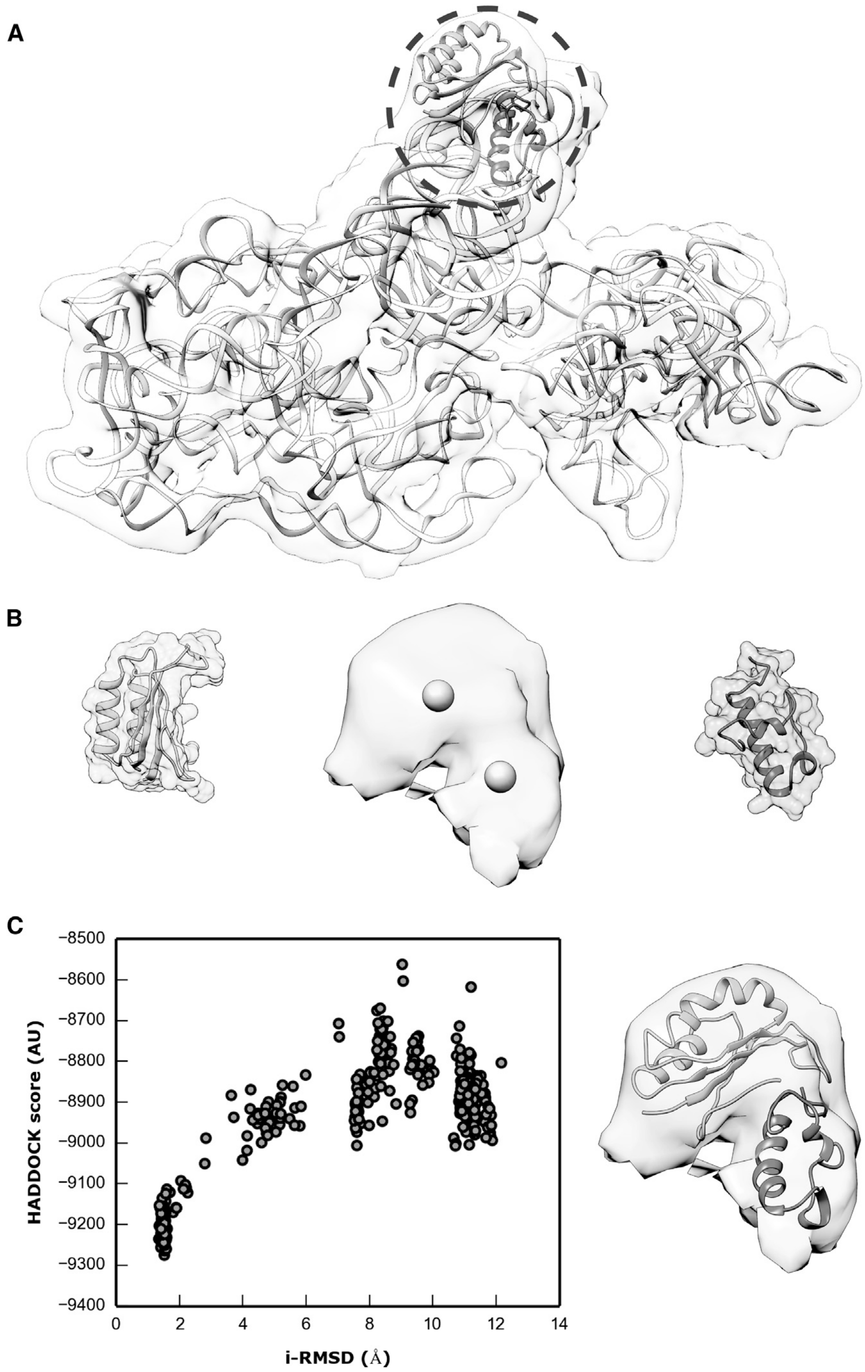


Figure 4. Cryo-EM Driven HADDOCKing Docking of the Ribosomal Proteins S7 and S19 onto the 30S *E. coli* Ribosome

(A) 30S *E. coli* ribosome structure (PDB: 2YKR) as it is currently fitted in the 9.8-Å cryo-EM map (EMD-1884). The ribosomal proteins S7 (light gray) and S19 (dark gray) and their corresponding density are circled.

(B) Docking setup used in HADDOCK showing the S7 and S19 protein, the centroids, and the density.

(C) The HADDOCK-EM score of the 400 refined models plotted versus their i-RMSD from the 2YKR structure. Next to it the solution with the best HADDOCK score and an i-RMSD of 1.56 Å is shown in the cryo-EM density.

hydroxyl radical footprinting data (Xu et al., 2008). Mutagenesis data show that the positively charged residues R221, R222, and K223 of KsgA are important in the interaction (Figure 5B) (Boehringer et al., 2012).

The 13.5-Å cryo-EM map has a corresponding current PDB model (4ADV). This model, however, contains a large number of clashes at the interface (>100). Furthermore, it reveals no favorable interactions and fails to give a clear explanation for the importance of the arginine residues identified by mutagenesis (Figure 5C). This is a typical side effect from manual rigid-body fitting. Running HADDOCK-EM using the radical footprinting, mutagenesis, and cryo-EM data results in a single cluster (Figure 6A) of which the best solution has an i-RMSD of 2.8 Å compared with the 4ADV model. The placement and orientations of the chains in the density are similar to the rigid-body fitted model as defined by the cryo-EM data. The HADDOCK-EM model is, however, of much better quality; it contains no clashes and reveals favorable hydrogen bonds made by R221, R222, and K223 with the backbone of the rRNA. Moreover, new potentially key residues can be identified, such as R147 and R248 (Figure 6B). Coincidentally, these newly identified residues are also highly conserved, corroborating our docking results (see Figure S3).

Modeling Virus-Antibody Complexes Using 8.5- and 21-Å Cryo-EM Data

To show the diverse range of systems that can be handled with HADDOCK, we applied our protocol on the adeno-associated virus 2 and immature Dengue virus complexed with antibodies for which 8.5- and 21-Å cryo-EM data and deposited models (3J1S and 3J42) are available, respectively.

For both cases, we performed a HADDOCK run combining the cryo-EM data with interface information. Since the binding regions on the antibody are known as well as the virus capsid proteins, residues that were within 5 Å of the other chain in the deposited atomic models were used as active residues. The solutions of the adeno-associated virus 2 converge into one cluster with an i-RMSD less than 1.5 Å from the deposited model (Figure 7A). However, when zooming in on the interface of the best-scoring HADDOCK model, the interactions show an extensive hydrogen bond network between the envelope protein and the antibody in contrast to the deposited model (Figure 7B).

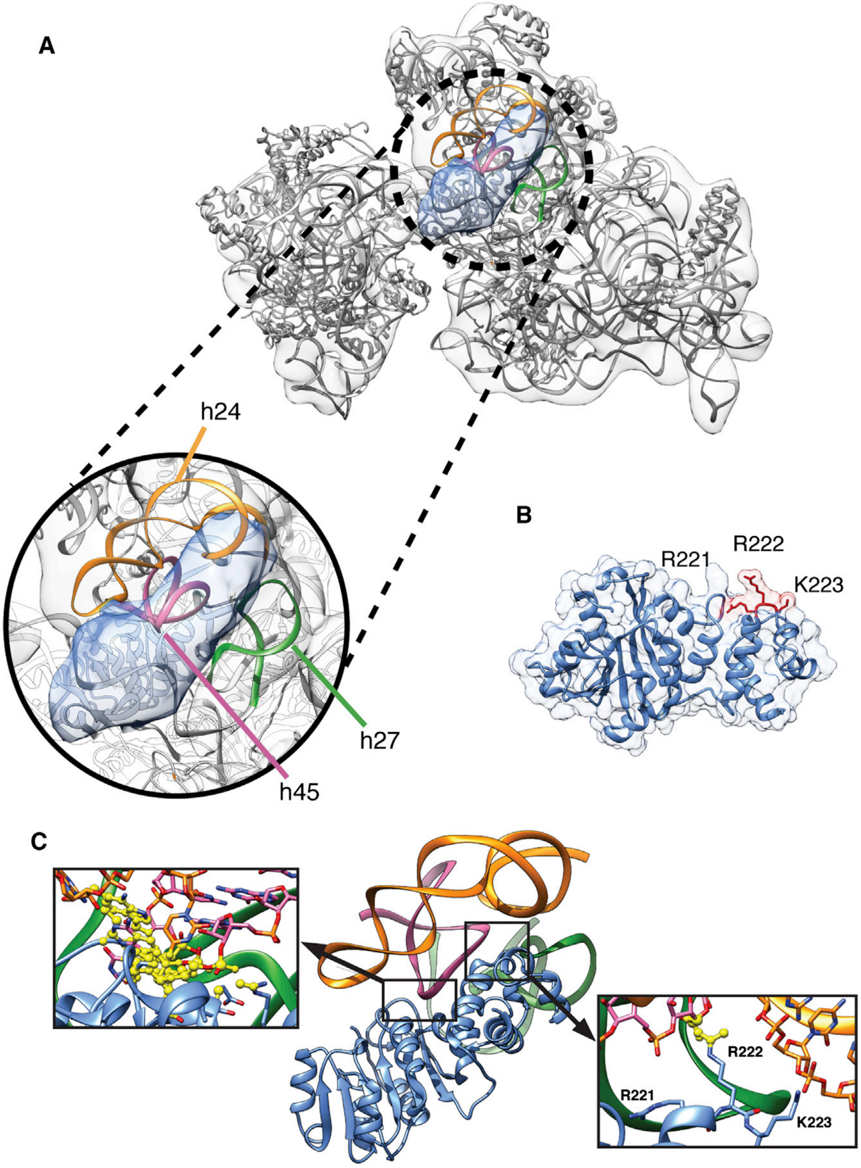


Figure 5. Cryo-EM and Mutagenesis Data of the 30S Maturing *E. coli* Ribosome and its Current Model

(A) The 13.5-Å cryo-EM map of the maturing 30S *E. coli* ribosome with its current PDB model fitted inside. The density of KsgA is shown in blue and the helices of the rRNA are shown in orange (h24), green (h27), and pink (h45). The binding site of KsgA is shown below enlarged.

(B) The crystal structure of KsgA of *E. coli* with the three key residues shown in red.

(C) A ribbon representation of the 4ADV model is shown in the middle. The left and right figures are close ups of the interface. Atoms displayed as yellow balls are clashes.

The HADDOCK solutions of the Dengue virus cluster into two groups, with an approximate i-RMSD of 2.0 and 4.5 Å with respect to the deposited model (Figure 7C). Inspecting the interface of the best-scoring HADDOCK model again shows favorable interactions between the prM protein of the Dengue virus with the antibody, while the 3J42 model lacks side chains and shows a backbone clash (Figure 7D).

Symmetrical Multibody Docking with Cryo-EM Data

HADDOCK is capable of using symmetry restraints to drive the docking of symmetrical assemblies. In order to combine symmetry and cryo-EM restraints, the rigid-body docking protocol was slightly modified compared with nonsymmetric complexes, with the main difference in the initial placement of the subunits (see Figure 8A; Experimental Procedures). We tested HADDOCK-EM with symmetry on the cyclic pentamer of the trypsin inhibitor (1B0C, Figure 8A). The ab initio mode of HADDOCK with C5 symmetry restraints results in two acceptable solutions after the refinement stage, with the best solution having an i-RMSD of 3.2 Å compared with the 1B0C structure. Adding cryo-EM data results in an increased number of acceptable solutions of 54, 400, and 400 when using 20-, 15-, and 10-Å resolution data, respectively, with the best models having i-RMSDs of 1.6, 1.0, and 0.7 Å (Figure 8B). Using higher-resolution data also results in more compact clusters, i.e. the distribution of i-RMSD values is reduced. When using 10-Å data only, a single near-native cluster is observed. At 20-Å resolution, multiple clusters appear and require the HADDOCK-EM score to discriminate the near-native cluster, which indeed has the best (lowest) HADDOCK-EM score.

We applied the symmetrical HADDOCK-EM protocol to model the pentameric large terminase complex of bacteriophage T7 using 16-Å negative stain EM data (EMD-2355, Daudén et al., 2013). As in the previous cases, the corresponding deposited model (4BIJ) shows clashes at the interfaces (Figure S8). The 400 generated HADDOCK models resulted in 33 clusters, with the best-scoring cluster having an i-RMSD of 2.9 Å compared with the 4BIJ model. Again, the interface of the best-scoring HADDOCK model alleviates the clashes and shows favorable interactions, while agreeing with the general binding mode of the 4BIJ model.

Conclusions

We have fully integrated cryo-EM data into HADDOCK, allowing the direct combination of cryo-EM data with all other available sources of information that HADDOCK supports, including symmetry and ambiguous interaction restraints. The performance of this integrative docking protocol was demonstrated using simulated cryo-EM data for a benchmark of 17 nonredundant protein-protein complexes: Including the cryo-EM data into the docking significantly increases both the quality and quantity of acceptable solutions, with higher-resolution data having a

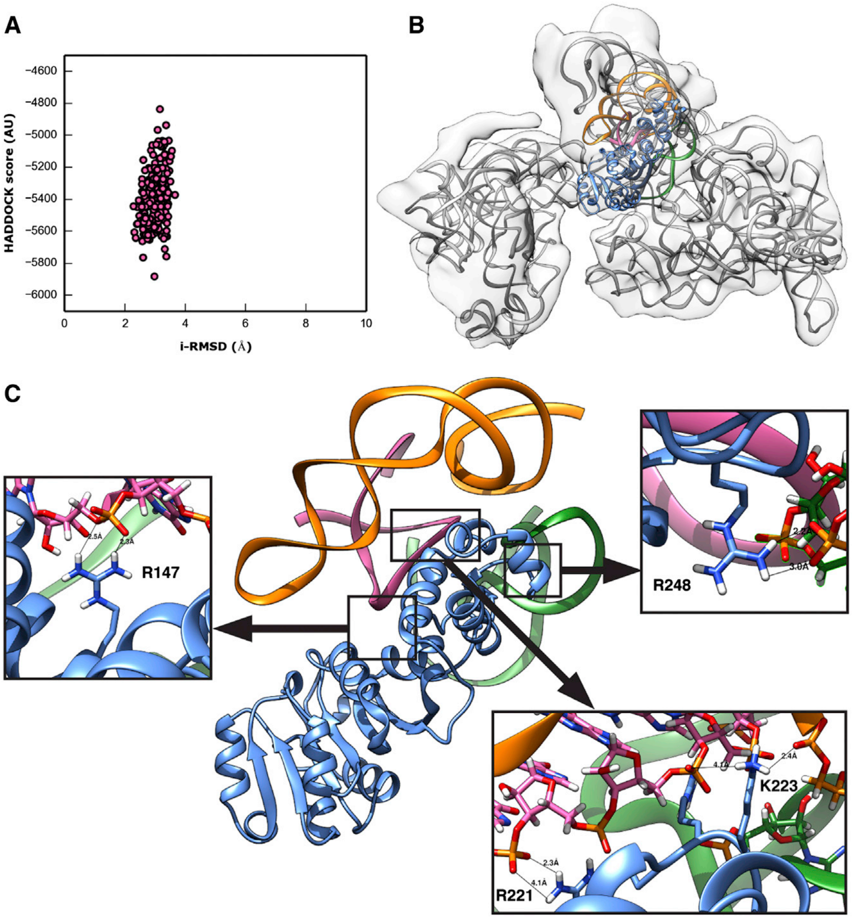


Figure 6. Cryo-EM Driven HADDOCKing of KsgA on Top of the 16S rRNA of *E. coli*

(A) The HADDOCK-EM score of the 400 refined models plotted versus the i-RMSD compared with the 4ADV model.

(B) Binding mode of the best-scoring HADDOCK-EM model, together with the 13.5-Å cryo-EM map.

(C) Close up of the binding of KsgA with the rRNA. The right bottom figure shows the favorable hydrogen bonds formed by the three key residues, R221, R222, and K223. At the left and upper right side, the additional evolutionary conserved residues R147 and R248 are shown forming favorable hydrogen bonds with the backbone of the rRNA.

larger impact. Its applicability was demonstrated on two ribosomes, two virus antibodies, and a symmetrical case using experimental data ranging from 8.5 to 21 Å. The integration of cryo-EM data with a proper physics-based force field and all other available information sources provides a powerful and user-friendly tool to generate high-quality, high-resolution models of macromolecular assemblies.

EXPERIMENTAL PROCEDURES

HADDOCK Protocol

HADDOCK has been described in details in previous work (Dominguez et al., 2003; De Vries et al., 2010b). Its docking protocol consists of three stages: an initial rigid-body docking stage (it0), a semi-flexible refinement stage using simulated annealing in torsion-angle space (it1), and a final flexible refinement stage in explicit water (itw). See the Supplemental Information for a detailed description of each stage.

After each stage, the models are scored with the following pseudo-energy functions:

$$E_{it0} = 0.1 \cdot E_{vdW} + 1.0 \cdot E_{elec} + 0.01 \cdot E_{AIR} + 1.0 \cdot E_{desolv} - 0.01 \cdot BSA \quad \text{Equation 1}$$

$$E_{it1} = 1.0 \cdot E_{vdW} + 0.2 \cdot E_{elec} + 0.1 \cdot E_{AIR} + 1.0 \cdot E_{desolv} - 0.01 \cdot BSA \quad \text{Equation 2}$$

$$E_{itw} = 1.0 \cdot E_{vdW} + 0.2 \cdot E_{elec} + 0.1 \cdot E_{AIR} + 1.0 \cdot E_{desolv} \quad \text{Equation 3}$$

where E_{it0}, E_{it1}, and E_{itw} are the scoring functions after the it0, it1, and itw stages, respectively, E_{vdW} is the intermolecular van der Waals energy, E_{elec} is the intermolecular electrostatic energy, E_{AIR} is the ambiguous interaction restraints energy, E_{desolv} is an empirical desolvation energy (Fernández-Recio et al., 2004), and *BSA* is the buried surface area in Å^2. The energies are calculated with an 8.5-Å cutoff based on OPLS (optimized potentials for liquid simulations) parameters (Jorgensen and Tirado-Rives, 1988).

HADDOCK-EM Protocol

As HADDOCK uses CNS (Crystallography and NMR System) (Brunger, 2007) as its computational engine, all crystallographic tools and energy functions available in CNS are available to HADDOCK. So the cryo-EM data, represented by a 3D real scalar field, can be directly read into the CNS framework and specific energy functions, typically in reciprocal space, can be used and applied. The HADDOCK-EM protocol uses in particular the xref energy term in CNS. It is very similar to the original HADDOCK method with some adjustments mainly in it0. A graphical representation of the adjusted it0 protocol is given in Figure 1. An integral part of our protocol is the use of centroids, where each centroid represents the approximate position of the COM of a subunit in the density map. When the resolution of the cryo-EM data decreases, the orientation of the subunits can be ambiguous but the approximate placement can still be determined. This is obvious in cases where several density maps are obtained with some subunits being alternately present and absent in the set, such as in the case of the ribosome (Xu et al., 2008). The position of the centroids can be determined in multiple ways. An objective way is to perform a full-exhaustive cross-correlation search to deduce regions of high cross-correlation values; the centroid can then be placed on the position with the highest value. They can be placed manually using graphics software; for example, UCSF Chimera has an option to place centroids in high-density regions in the map. Another option is to place an atomic structure in the density at an approximately correct position, calculate its COM, and use this as the position of the centroid. Methods for automatic simultaneous detection of centroids have also been reported (Birmanns and Wriggers, 2007; Lasker et al., 2010; Wriggers et al., 1998; Zhang et al., 2010). A more elaborate approach uses experimental data in conjunction with the cryo-EM map to infer the positions of the subunits, as was shown for RNA polymerase II (Murakami et al., 2013). The centroids are entered into HADDOCK-EM as Cartesian coordinates in the start parameters. Together with the cryo-EM map and its resolution, they represent all the input required to run HADDOCK-EM.

During the docking, each subunit is given a random orientation and initially placed on a sphere centered on the midpoint of the centroids. In the case of two chains, the subunits are placed opposite each other on the sphere with a minimal distance of 25 Å between them. Afterward, for each docking trial, they are given a random rotation and translation within a 10-Å box to enhance the sampling. Distance restraints are defined between the COM of each subunit and either all determined centroids are ambiguous restraints in cases where the placement of the subunit in the density is ambiguous or a

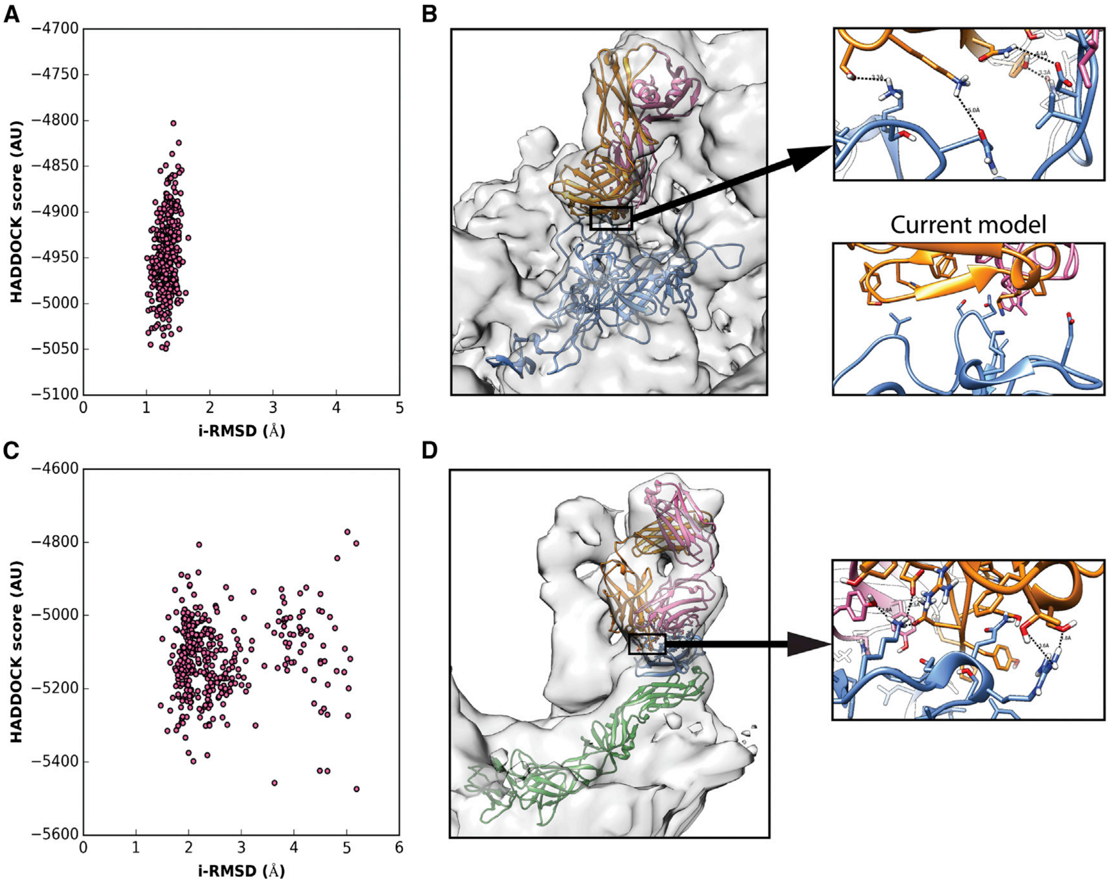


Figure 7. Virus-Antibody HADDOCKing Using 8.5- and 21-Å Cryo-EM Data

(A) The HADDOCK-EM score of the 400 generated models of the adeno-associated virus 2-antibody complex versus their i-RMSD using 3J1S as a reference.

(B) Best-scoring HADDOCK model shown in the cryo-EM density. The envelope protein (blue) forms favorable interactions with the antibody A20 chains (orange and pink). The 3J1S interface is shown under the interface close up.

(C) The HADDOCK-EM score of the 400 generated models of the Dengue virus-antibody complex versus their i-RMSD using 3J42 as a reference results in two clusters.

(D) Best-scoring HADDOCK model shown in the cryo-EM density. The Dengue envelope protein (green) with the prM protein (blue) forms favorable interactions with the 2H2 Fab-fragment (orange and pink).

specific centroid if the placement is unambiguous. The distance restraint is described by a soft square potential between two pseudo-atoms, one of which corresponds to the centroid and the other to the COM of the subunit. An initial complex is formed by rigid-body energy minimization, where the energy is a combination of the force field, the centroid-based distance restraints, and other possible experimentally based distance and orientation restraints.

After the initial energy minimization, for binary systems we properly orient the complex in the density by performing a fine full-exhaustive search around the axis that is formed by the line joining the centroids in 4° increments. Each orientation is scored by the vector residual energy term in CNS, given by

$$E_{vector} = \frac{\sum_{H} (F_{em} - F_c)^2}{\sum_{H} {F_{em}}^2} \qquad \text{Equation 4}$$

where the summation is over all the Miller indices H up to the specified resolution of the cryo-EM map, and F_{em} and F_c are the complex-valued Fourier coefficients of the cryo-EM map and the calculated density, respectively. It should be noted that minimizing the vector residual in reciprocal space is mathematically the same as maximizing the cross-correlation in real space (Navaza et al., 2002) and thus we refer to this potential simply as the cross-correlation. The complex is reoriented in the density conforming to the optimal cross-correlation value found during the search. A final rigid-body energy minimization is performed directly against the map using the cross-correlation (vector potential energy term in CNS), van der Waals, and electrostatic energy terms. For each complex typically 10,000 models are generated this way.

The models are then scored by adding an LCC term to the regular HADDOCK score (Equations 1–3), where the LCC is given by

$$LCC = \frac{\sum_{i} \left(\rho_{em} - \overline{\rho_{em}}\right) \cdot \left(\rho_c - \overline{\rho_c}\right)}{\sigma_{em}\sigma_c} \qquad \text{Equation 5}$$

where the summation is over the voxels i which are maximally 3 Å away from an atom of the model, ρ_{em} is the density value at voxel i of the cryo-EM map, $\overline{\rho_{em}}$ is the average density value of all the voxels i, ρ_c is the density value at voxel i of the calculated density, $\overline{\rho_c}$ is the average density value of all the voxels i of the calculated density and σ_{em} and σ_c are the SDs of the cryo-EM and calculated density over the voxels i, respectively. The HADDOCK-EM scores are thus given by

$$\begin{aligned} E_{it0,EM} = {} & 0.1 \cdot E_{vdW} + 1.0 \cdot E_{elec} + 0.01 \cdot E_{AIR} \\ & + 1.0 \cdot E_{desolv} - 0.01 \cdot E_{BSA} - w_{it0} \cdot LCC \end{aligned} \qquad \text{Equation 6}$$

$$E_{itw,EM} = 1.0 \cdot E_{vdW} + 0.2 \cdot E_{elec} + 0.1 \cdot E_{AIR} + 1.0 \cdot E_{desolv} - w_{itw} \cdot LCC \qquad \text{Equation 7}$$

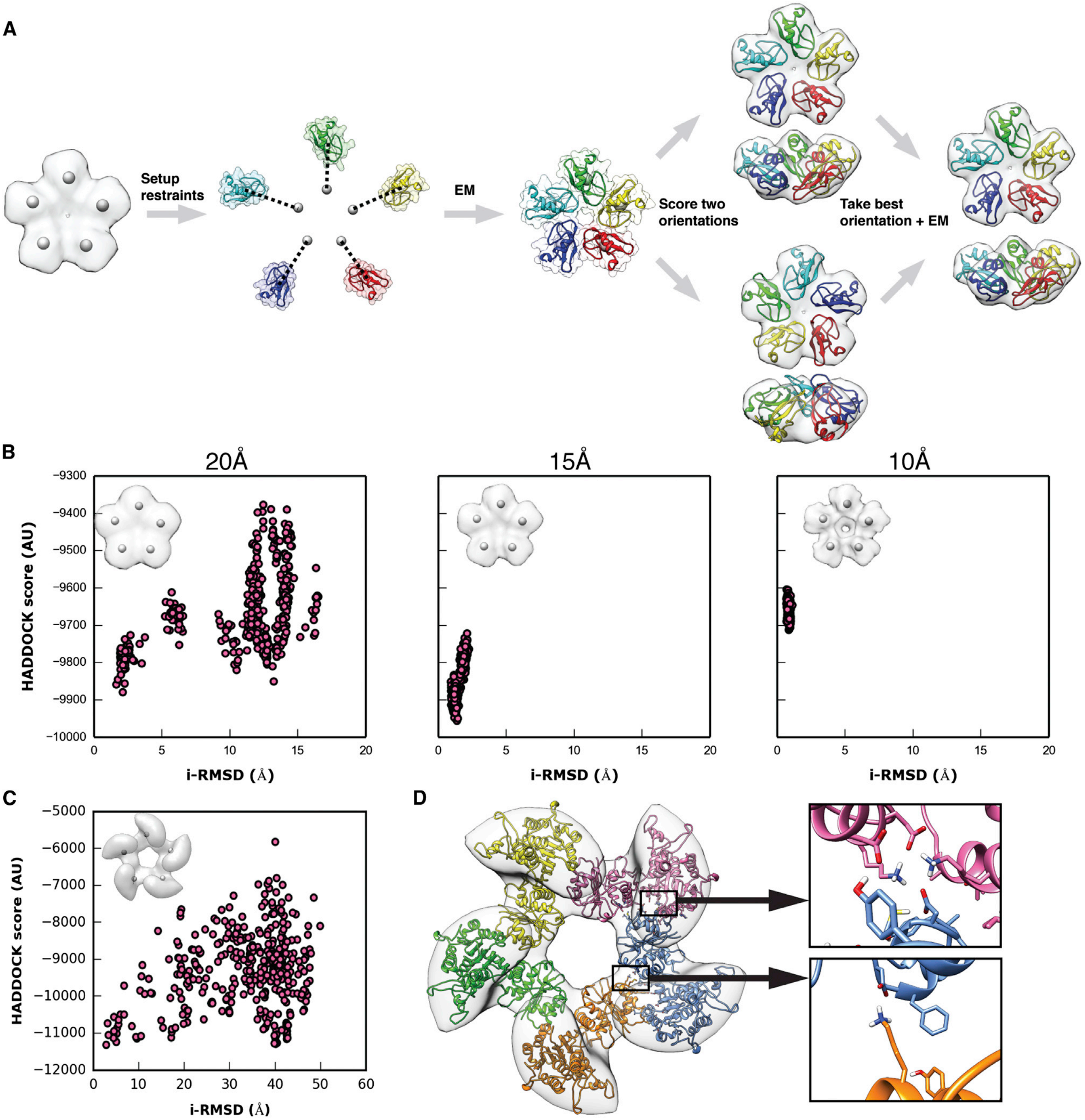


Figure 8. HADDOCK-EM with Symmetry Protocol Applied on the Trypsin Inhibitor and Large Terminase Pentamer

(A) Protocol of HADDOCK-EM with symmetry during the rigid-body stage. After determining the centroids in the density, each subunit is placed on a circle with its center on the midpoint of the centroids. C5 symmetry is imposed on the system from the beginning and ambiguous distance restraints are generated between the COM of each subunit and each centroid. An initial complex is formed by a rigid-body energy minimization (EM). To orient the complex properly in the density, we calculate the cross-correlation of two orientations of the complex with the cryo-EM data. A second round of rigid-body energy minimization is performed on the orientation with the highest cross-correlation directly against the cryo-EM data.

(B) The HADDOCK-EM score versus the i-RMSD compared with the native complex (1B0C) are plotted for the 400 refinement complexes using 20-, 15-, and 10-Å simulated data.

(C) The HADDOCK-EM score of the 400 generated large terminase models versus the i-RMSD. The 16-Å negative stain data together with the centroids is shown in the left corner.

(D) Best-scoring HADDOCK model shown in the cryo-EM density. Two close ups of the interface are displayed to the right of it.

where w_{it0} and w_{itw} are weight terms for the LCC pseudo-energy that need to be determined (see below). The top 400 best-scoring structures are selected for further flexible refinement in it1 and itw. The refinement protocols are similar to the standard HADDOCK protocol, however the energy now also contains the additional cross-correlation-based energy term in addition to the other force field and restraint energy terms.

It should be noted that the maximum number of subunits that can be docked simultaneously is currently restricted to six (this limitation will be lifted in a future version). Furthermore, in order to use the HADDOCK-EM protocol, approximate knowledge of the position of each subunit in the form of centroids is a requisite for a successful docking run. Other minor requirements are that the number of voxels in each dimension of the cryo-EM data is a multiple of two, three, and five to calculate the fast Fourier transforms (FFTs) used in the cross-correlation potential, and that the density should be converted to CNS/XPLOR format. For the latter two tasks, Python scripts are included in the HADDOCK distribution. Finally, the time required for a HADDOCK-EM run decreases with decreasing map size, since this speeds up the calculations of the FFTs.

Optimizing and Benchmarking HADDOCK-EM

The HADDOCK-EM protocol relies on the optimization of two parameters for the docking, namely the force constant for the centroid-based distance restraints and the weight for the cross-correlation energy term. In addition, the weight factors of the LCC term in the it0 and itw HADDOCK score need to be determined. For this, we used a benchmark consisting of 17 complexes taken from the protein-protein docking benchmark 4.0 (Hwang et al., 2010) (see Table 1). Centroids were determined by calculating the COM of each unbound chain that is optimally superimposed onto the native complex. Simulated cryo-EM data were calculated using a Python script, based on the *molmap* function in UCSF Chimera (see Supplemental Information).

We first determined the centroid-based force constant by running the benchmark at different values for the force constant, creating 10,000 models for each complex in it0. Since the force constant is only used in it0, the structures were not scored or refined. The value for the force constant that gave the most acceptable solutions, where an acceptable solution is defined as having an i-RMSD $\leq$4.0 Å compared with the native complex, was chosen (results not shown). The i-RMSDs were calculated using ProFit (Martin, A.C.R., http://www.bioinf.org.uk/software/profit/). For the determination of the weight factor for the cross-correlation term, we followed the same protocol but with the optimized force constant for the centroid-based distance restraints using simulated data at 10, 15, and 20 Å, which were generated as described above. This gave a value of 50 for the force constant and a weight factor of 15,000 for the cross-correlation-based energy term, independent of the resolution (results not shown).

The weight factor for the LCC in the it0-HADDOCK-EM score was determined by running the benchmark using the optimized parameters and varying the LCC weight in order to maximize the number of acceptable solutions in the top 400 at the three resolutions of 10, 15, and 20 Å. This gave a value of $-$400. The LCC weight factor in the itw score was determined by maximizing the number of acceptable solutions in the top 20, which gave a weight factor of $-$10,000.

To investigate the sensitivity of the protocol to incorrectly placed centroids, we ran five cases of the benchmark with displaced centroids. Each centroid was moved in a random direction by taking a random point on the unit sphere with a displacement of 3, 5, and 7 Å. The solutions were analyzed as explained above.

HADDOCK-EM with Symmetry

To leverage C_n symmetry in cyclical symmetric complexes, a few adjustments were made to the nonsymmetric HADDOCK-EM protocol (Figure 8A). The main difference is in the initial placement of the subunits in the it0 stage. Instead of placing the subunits on a sphere, we place them on a circle with its center placed on the middle point of the centroids and parallel to the plane of the centroids. The radius of the circle is chosen such that the minimal distance between two subunits is at least 25 Å. The requested C_n symmetry is imposed on the system from the start.

After the initial placement, ambiguous centroid-based distance restraints are generated, i.e. we create a distance restraint between the COM of each subunit to each centroid. We form an initial complex again by performing rigid-body energy minimization, where the energy includes the force field, the centroid-based distance restraints, and the available symmetry restraints already in HADDOCK. Once the initial complex is formed, it needs to be properly oriented in the density. Only two orientations need to be sampled for this, namely the current orientation and the upside-down complex. The orientation corresponding to the highest cross-correlation with the cryo-EM data is chosen. A final rigid-body energy minimization is performed against the map, using the cross-correlation potential in combination with the force field and symmetry restraints. Typically, 10,000 models are generated. They are scored with the it0-HADDOCK-EM and 400 models are refined in the it1 and itw stages. The refinement protocol is similar to the nonsymmetric HADDOCK-EM protocol, but with added symmetry restraints.

Modeling Complexes Using Experimental Data

All high-resolution models were downloaded from the PDB and the cryo-EM data were from the EMDataBank. Details about the determination of the centroids and setup of the docking run for each specific case can be found in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, eight figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.str.2015.03.014.

ACKNOWLEDGMENTS

Financial support from the Dutch Foundation for Scientific Research (NWO) (ECHO grant no.711.011.009) is acknowledged.

Received: September 26, 2014
Revised: March 12, 2015
Accepted: March 13, 2015
Published: April 23, 2015

REFERENCES

Alber, F., Förster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem. *77*, 443–477.

Baker, T.S., and Johnson, J.E. (1996). Low resolution meets high: towards a resolution continuum from cells to atoms. Curr. Opin. Struct. Biol. *6*, 585–594.

Birmanns, S., and Wriggers, W. (2007). Multi-resolution anchor-point registration of biomolecular assemblies and their components. J. Struct. Biol. *157*, 271–280.

Boehringer, D., O'Farrell, H.C., Rife, J.P., and Ban, N. (2012). Structural insights into methyltransferase KsgA function in 30S ribosomal subunit biogenesis. J. Biol. Chem. *287*, 10453–10459.

Brunger, A.T. (2007). Version 1.2 of the Crystallography and NMR system. Nat. Protoc. *2*, 2728–2733.

Daudén, M.I., Martín-Benito, J., Sánchez-Ferrero, J.C., Pulido-Cid, C., Valpuesta, J.M., and Carrascosa, J.L. (2013). Large terminase conformational change induced by connector binding in bacteriophage T7. J. Biol. Chem. *288*, 16998–17007.

De Vries, S.J., and Zacharias, M. (2012). ATTRACT-EM: a new method for the computational assembly of large molecular machines using cryo-EM maps. PLoS One *7*, e49733.

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

De Vries, S.J., Melquiond, A.S.J., Kastritis, P.L., Karaca, E., Bordogna, A., van Dijk, M., Rodrigues, J.P.G.L.M., and Bonvin, A.M.J.J. (2010b). Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. Proteins *78*, 3242–3249.

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.

Esquivel-Rodríguez, J., and Kihara, D. (2012). Fitting multimeric protein complexes into electron microscopy maps using 3D Zernike descriptors. J. Phys. Chem. B *116*, 6854–6861.

Esquivel-Rodríguez, J., and Kihara, D. (2013). Computational methods for constructing protein structure models from 3D electron microscopy maps. J. Struct. Biol. *184*, 93–102.

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

Goddard, T.D., Huang, C.C., and Ferrin, T.E. (2007). Visualizing density maps with UCSF Chimera. J. Struct. Biol. *157*, 281–287.

Guo, Q., Yuan, Y., Xu, Y., Feng, B., Liu, L., Chen, K., Sun, M., Yang, Z., Lei, J., and Gao, N. (2011). Structural basis for the function of a small GTPase RsgA on the 30S ribosomal subunit maturation revealed by cryoelectron microscopy. Proc. Natl. Acad. Sci. USA *108*, 13100–13105.

Hoang, T.V., Cavin, X., and Ritchi, D.W. (2013). gEMfitter: a highly parallel FFT-based 3D density fitting tool with GPU texture memory acceleration. J. Struct. Biol. *184*, 348–354.

Huang, S.-Y. (2014). Search strategies and evaluation in protein-protein docking: principles, advances and challenges. Drug Discov. Today *19*, 1081–1096.

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein-protein docking benchmark version 4.0. Proteins *78*, 3111–3114.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J.E., Vajda, S., Vakser, I., and Wodak, S.J. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. Proteins *52*, 2–9.

Joerger, A.C., and Fersht, A.R. (2007). Structure-function-rescue: the diverse nature of common p53 cancer mutants. Oncogene *26*, 2226–2242.

Jorgensen, W.L., and Tirado-Rives, J. (1988). The OPLS potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J. Am. Chem. Soc. *110*, 1657–1666.

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

Karaca, E., and Bonvin, A.M.J.J. (2013a). Advances in integrative modeling of biomolecular complexes. Methods *59*, 372–381.

Karaca, E., and Bonvin, A.M.J.J. (2013b). On the usefulness of ion-mobility mass spectrometry and SAXS data in scoring docking decoys. Acta Crystallogr. D Biol. Crystallogr. *69*, 683–694.

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.

Kawabata, T. (2008). Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. Biophys. J. *95*, 4643–4658.

Lage, K. (2014). Protein-protein interactions and genetic diseases: the interactome. Biochim. Biophys. Acta *1842*, 1971–1980.

Lasker, K., Sali, A., and Wolfson, H.J. (2010). Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. Proteins *78*, 3205–3211.

McCraw, D.M., O'Donnel, J.K., Taylor, K.A., Stagg, S.M., and Chapman, M.S. (2012). Structure of adeno-associated virus-2 in complex with neutralizing monoclonal antibody A20. Virology *431*, 40–49.

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

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nat. Methods *10*, 47–53.

Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D.A., Adams, C.M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B.J., et al. (2013). Architecture of an RNA polymerase II transcription pre-initiation complex. Science *342*, 1238724.

Navaza, J., Lepault, J., Rey, F.A., Alvarez-Rúa, C., and Borge, J. (2002). On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. Acta Crystallogr. D Biol. Crystallogr. *58*, 1820–1825.

Nero, T.L., Morton, C.J., Holien, J.K., Wielens, J., and Parker, M.W. (2014). Oncogenic protein interfaces: small molecules, bit challenges. Nat. Rev. Cancer *14*, 248–262.

Orlova, E.V., and Saibil, H.R. (2011). Structural analysis of macromolecular assemblies by electron microscopy. Chem. Rev. *111*, 7710–7748.

Petrey, D., and Honig, B. (2014). Structural bioinformatics of the interactome. Annu. Rev. Biophys. *43*, 193–210.

Rodrigues, J.P.G.L.M., and Bonvin, A.M.J.J. (2014). Integrative computational modeling of protein interactions. FEBS J. *281*, 1988–2003.

Schneidman-Duhovny, D., Rossi, A., Avila-Sakar, A., Kim, S.J., Velázquez-Muriel, J., Strop, P., Liang, H., Kurkenberg, K.A., Liao, M., Kim, H.M., et al. (2012). A method for integrative structure determination of protein-protein complexes. Bioinformatics *28*, 3282–3289.

Shacham, E., Sheehan, B., and Volkmann, N. (2007). Density-based score for selecting near-native atomic models of unknown structures. J. Struct. Biol. *158*, 188–195.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. Structure *16*, 295–307.

Trellet, M., Melquiond, A.S.J., and Bonvin, A.M.J.J. (2013). A unified conformational selection and induced fit approach to protein-peptide docking. PLoS One *8*, e58769.

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

Velázquez-Muriel, J., Lasker, K., Russel, D., Phillips, J., Webb, B.M., Schneidman-Duhovny, D., and Sali, A. (2012). Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images. Proc. Natl. Acad. Sci. USA *109*, 18821–18826.

Wang, Z., Li, L., Pennington, J.G., Sheng, J., Yap, M.L., Plevka, P., Meng, G., Sun, L., Jiang, W., and Rossmann, M.G. (2013). Obstruction of dengue virus maturation by Fab fragments of the 2H2 antibody. J. Virol. *87*, 8909–8915.

Wells, R., and McClendon, C.L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature *450*, 1001–1009.

Wriggers, W., Milligan, R.A., Schulten, K., and McCammon, J.A. (1998). Self-organizing neural networks bridge the biomolecular resolution gap. J. Mol. Biol. *284*, 1247–1254.

Xu, Z., O'Farrell, H.C., Rife, J.P., and Culver, G.M. (2008). A conserved rRNA methyltransferase regulates ribosome biogenesis. Nat. Struct. Mol. Biol. *15*, 534–536.

Zhang, S., Vasishtan, D., Xu, M., Topf, M., and Alber, F. (2010). A fast mathematical programming procedure for simultaneous fitting of assembly components into cryo-EM density maps. Bioinformatics *26*, i261–i268.