Chapter 10

Information-Driven Modeling of Protein-Peptide Complexes

Mikael Trellet, Adrien S.J. Melquiond, and Alexandre M.J.J. Bonvin

Abstract

Despite their biological importance in many regulatory processes, protein-peptide recognition mechanisms are difficult to study experimentally at the structural level because of the inherent flexibility of peptides and the often transient interactions on which they rely. Complementary methods like biomolecular docking are therefore required. The prediction of the three-dimensional structure of protein-peptide complexes raises unique challenges for computational algorithms, as exemplified by the recent introduction of protein-peptide targets in the blind international experiment CAPRI (Critical Assessment of PRedicted Interactions). Conventional protein-protein docking approaches are often struggling with the high flexibility of peptides whose short sizes impede protocols and scoring functions developed for larger interfaces. On the other side, protein-small ligand docking methods are unable to cope with the larger number of degrees of freedom in peptides compared to small molecules and the typically reduced available information to define the binding site. In this chapter, we describe a protocol to model protein-peptide complexes using the HADDOCK web server, working through a test case to illustrate every steps. The flexibility challenge that peptides represent is dealt with by combining elements of conformational selection and induced fit molecular recognition theories.

Key words Biomolecular interactions, Information-driven docking, Conformational changes, Flexibility, HADDOCK, Molecular modeling

1 Introduction

A large variety of methods are available to scientists to investigate the 3D structure of biomolecular complexes. Experimental determination of protein-peptide complexes is, however, often nontrivial due to the dynamic nature of the transient interactions they mediate. While X-ray crystallography is struggling with the high flexibility of peptides, hybrid approaches that rely on an experimental characterization of the binding site (NMR, cross-linking mass spectrometry …) and/or NMR-derived restraints to limit the conformational space of the peptide (e.g. dihedral angle restraints), in combination with computational modeling, have demonstrated

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_10,

their accuracy for various protein-peptide systems [1–6]. Structural characterisation of low affinity interactions remain unfortunately out of reach for most experimental methods. There is therefore a need for improving existing computational methods.

Modeling of protein-protein complexes has a long-standing history that started back in the late 1970s with the first automated computer analysis of protein-protein interactions [7]. Macromolecular docking made its first proof of concept with the successful prediction of the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase in 1996 [8]. Protein-peptide interactions, in contrast, have only been studied computationally recently. The recognition mechanisms underlying their assembly are still debated [9–12]. Flexibility is a key characteristic of peptides, which are short polypeptidic chains ranging from 5 to 30 amino acids and, in most cases, do not adopt a well-defined conformation when unbound, i.e., in their free state. This represents a major challenge for classical docking algorithms where both constituents are usually treated as rigid in first instance, to be refined at later stages, allowing some degrees of flexibility at the interface.

Over the last years, a number of new algorithms or adaptations of existing docking methods have been released to address the unique challenges raises by protein-peptide interactions [13–21]. Based on the HADDOCK framework [22], we have developed an original approach that combines ensemble docking and enhanced flexibility to improve the sampling of peptides [23]. HADDOCK is an information-driven docking software [24] using CNS (Crystallography and NMR system) [25, 26] as computational engine and the OPLS united atom force field [27] to calculate the non bonded interactions (with a cutoff of 8.5 Å). It allows the integration of a variety of experimental data to drive the docking process, such as NMR chemical shift perturbation and mutagenesis data. HADDOCK also introduces flexibility into the subunits during the docking process, ending with a final refinement of the models in explicit solvent. Currently, HADDOCK is one of the most cited docking software [28], counts a large community of 3,700+ users worldwide, and ranks among the best performing docking methods based on CAPRI (Critical Assessment of PRediction of Interactions) [29], a community wide experiment where participants have a limited time to predict the structure of a complex given only the structures, sometimes even only the sequences, of its free constituents.

We have recently optimized HADDOCK's protocol for protein-peptide docking against a benchmark of 101 protein-peptide complex structures, achieving a remarkable overall performance when starting from unbound structures [23]. In this chapter, we describe step by step this protocol using the HADDOCK web server.

2 Theory

This section describes the different steps and their background in order to perform a protein-peptide docking run and achieve the overall best performances with HADDOCK.

2.1 Peptide Conformation Sampling

Unlike protein-protein docking, we usually do not have access either to the free form of the peptide or to any structural template that could be used to generate a reliable starting 3D model of the structure of the peptide. To solve this problem, we have proposed a specific protocol for flexible docking of short peptides (5–15 amino acids) that starts from an ensemble of three different conformations of the peptide (α-helix, polyproline-II, and extended—*see* Subheading 3 for more details about how to generate this ensemble). This canonical ensemble does not aim at discretizing the conformational space sampled by the free peptide, but rather represents conformations often observed in protein-peptide complexes. Indeed, taken together, these three conformations cover about 80 % of the observed peptide-bound structures in the Protein Data Base [30]. Building onto the ensemble docking capability of HADDOCK, protein-peptide docking can start from these three distinct conformations and, hopefully, select be best suited peptide conformation for the complex under study, following a conformational selection mechanism.

2.2 Interface Restraints

HADDOCK uses ambiguous and unambiguous restraints throughout the entire docking process to drive the complex formation (*see* Subheading 2.3 for more details). These restraints can be derived from various experimental information sources such as NMR chemical shifts perturbations, hydrogen/deuterium exchange, chemical cross-linking detected by mass spectrometry, mutagenesis … [31, 32]. All this information is usually translated into distance or angle restraints used both for sampling and scoring. In this protocol we describe a classical scenario in which no information is available about which residues of the peptide are involved in binding, treating it as fully "passive," which means peptide residues can make contacts but no penalty will be paid if they do not. One the protein side we define a large surface centred on the native interface. For each docking trial, we randomly select half of the so-called active residues that belong to this surface (making the assumption that they are directly involved in the binding) and define ambiguous restraints toward the peptide.

2.3 Protein-Peptide HADDOCKing

2.3.1 Docking Protocol

The docking protocol in HADDOCK consists of three successive steps:

- *it0*: Rigid-body energy minimization (RBEM)
- *it1*: Semiflexible simulated annealing (SA) in torsion angle space (TAD/SA)
- *Water*: Final restrained molecular dynamics in explicit solvent

Pre- and post-processing steps are performed: (1) to build missing atoms in the preliminary step and (2) to launch energetic, intermolecular, and restraint analyses in the final step. For further details please refer to [22, 24].

One critical aspect in protein-peptide recognition is the importance of long-range electrostatic interactions [17]. Therefore, the user should specify charged Cter and Nter (default in HADDOCK) when working with naturally occurring peptides or uncharged termini when the peptide is a fragment of protein or capped in the experiment, this to avoid undesired interaction with the termini in the latter case (*see* **Note 1**).

Rigid-Body Energy Minimization (RBEM, it0)

In this initial docking stage, the interacting partners are first separated in space and randomly rotated around their respective center of mass. As a result, the starting positions of peptides adopt a spherical distribution around the protein receptor. The number of models generated in this step should typically be increased from the default 1,000–6,000, to ensure that each of the three distinct peptide conformations from the canonical ensemble is sampled 2,000 times. The resulting models are ranked according to the HADDOCK score (*see* below), and the top ranking models (here the top 400) are selected for further flexible refinement.

Semiflexible Simulated Annealing in Torsion Angle Space (TAD/SA, it1)

Four stages of SA are performed in *it1* influencing, respectively, the orientation of the components, the side chains at the interface, and finally both side chains and backbone of the interface residues. This semiflexible refinement stage is quite crucial in protein-peptide binding since it allows the peptide to fold and adapt its conformation to the protein binding site. To maximize the chance of finding a correct conformation at this stage, the peptide is treated as fully flexible over all four stages of the simulated annealing refinement. The protein is treated as default, with its interface residues becoming flexible in the last two stages. Further, we increase the number of simulation steps by a factor 4 for the successive stages of the simulated annealing refinement (from the default 500/500/1,000/1,000 to 2,000/2,000/4,000/4,000) to increase sampling. In order to avoid deformation of helical models that may have been selected after it0, dihedral angle restraints are applied to these (*see* **Note 2**).

Restrained Molecular Dynamics in Explicit Solvent (Water)

The structures obtained after simulated annealing are finally refined in an explicit solvent layer to further improve their scoring. This is done by molecular dynamics simulation in water, solvating the complex in an 8 Å shell of TIP3P water molecules [33].

2.3.2 Clustering of Final Solutions

The final models generated by HADDOCK are clustered based on their interface-RMSD using a 5 Å cutoff instead of the 7.5 Å cutoff used for clustering protein-protein poses (*see* **Note 3**).

A smaller value is required in order to ensure conformational homogeneity of the clusters due to the smaller size of the peptides compared to full proteins.

2.3.3 Quality Criteria

To assess the quality of the generated models, we follow the CAPRI standards [34, 35]. We will use mainly the interface-RMSD (i-RMSD), which is calculated on backbone atoms of both protein and peptide residues which are within 10 Å from each other in the reference crystal structures of the complex. The calculation of i-RMSD between a model and a reference is done in two steps that are illustrated in Fig. 1:

1. We fit the protein of the model onto the protein of the reference.
2. We calculate the positional root-mean-square deviation between the model and the reference structures for the backbone atoms of the interface residues (protein + peptide).

To account for the small size of peptides, the standard CAPRI acceptability thresholds need to be decreased:

- Not acceptable: i-RMSD > 2 Å.
- Near-native prediction: 1 Å ≤ i-RMSD ≤ 2 Å.
- High-quality (subangstrom) prediction: i-RMSD < 1 Å.

3 Methods

In order to successfully run this protocol using the HADDOCK web server, two software programs need to be installed locally. First, the input ensemble of three conformations for the peptide can be generated using the PyMOL script provided in the supplementary material associated with this chapter from the *Springer extra* web site (http://extras.springer.com). PyMOL [36] is a molecular visualization system, free for educational use (http://www.pymol.org). Secondly, the models are compared based on RMSD values calculated using ProFit, a free program for protein structure least squares fitting (http://www.bioinf.org.uk/software/profit/index.html). Finally, a web browser, an internet connexion and registration to the HADDOCK web server are the only pre-requisites to access the HADDOCK web server.

In the following sections, we illustrate our protocol on a test case taken from the benchmark dataset [11]. The protocol should be run on a GNU/Linux system or under Mac OSX.

3.1 Modeling of Complexes with HADDOCK

In this section, we model the peptide `DAIDALSSDFT`, corresponding to the disordered region of the calpastatin inhibitory domain C, in complex with the calpain domain VI, a proteolytic enzyme

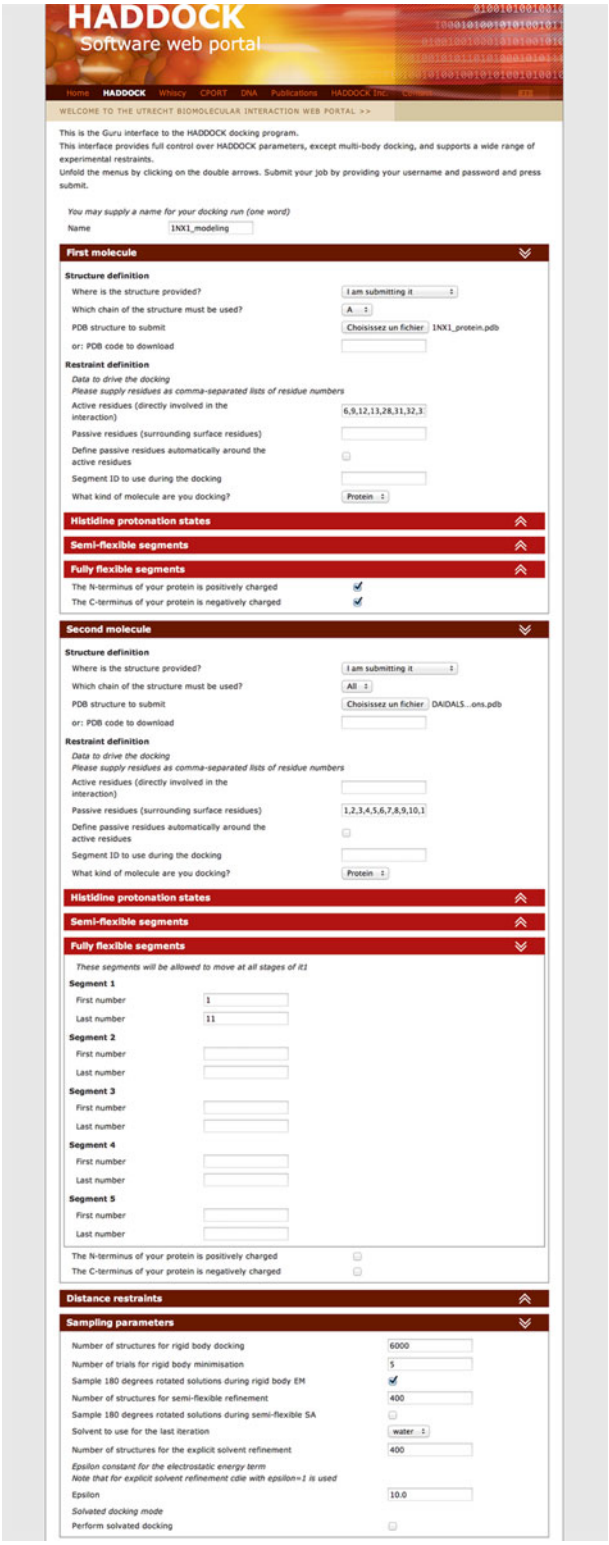


Fig. 1 Overview of the HADDOCK web server Guru interface (accessible from http://haddock.science.uu.nl/services/HADDOCK). A click on the *right arrows* will expand the associated sections to display HADDOCK parameters and/or input fields. In the current view, the *First molecule* and *Second molecule* and *Sampling parameters* sections are expanded. Fields are filled with necessary input for the docking example provided in Subheading 3.1.2

involved in a number of cell functions such as cell mobility and cell cycle progression. The coordinates of both the complex (PDBid: 1NX1) and the unbound structure of the calpain domain VI (PDBid: 1ALV) are available.

3.1.1 Preparation of PDB Files

Each PDB provided to HADDOCK has to respect the PDB format with proper syntax and clear chain identifiers (*see* **Note 4**). The input ensemble for the peptide will be composed of three artificially generated models using PyMOL [36]. Each model corresponds to a specific conformation of the peptide we want to dock onto its associated protein receptor. A PyMol script adapted from an original script of Robert L. Campbell (http://pldserver1.biochem.queensu.ca/~rlc/work/pymol/) is provided to facilitate the creation of the ensemble.

1. Open PyMol and execute the script to access its functions, in the PyMol console, type:

   ```
   > run 3c_build_seq.py
   ```

2. Use the building function provided by the script. For instance, to create the three conformations of the calpastatin peptide required to start the docking run, we type in PyMol:

   ```
   > build_seq extended_pept, DAIDALSSDFT, ss=extended
   > build_seq helical_pept, DAIDALSSDFT, ss=helix
   > build_seq polypro_pept, DAIDALSSDFT, ss=polypro
   ```

3. You can now save the structure coordinates in the PDB format via the Menu File->Save Molecule…
4. Once the three conformations (extended/helix/polyproline II) have been built and saved, the corresponding PDB files have to be merged into a unique PDB file before we can use them as input in HADDOCK. Each conformation must be defined as a unique MODEL, just alike NMR ensemble, meaning that the coordinates of each model must start with a MODEL statement and end with an ENDMDL statements in the PDB coordinate file. This can easily be done with a simple text editor.

The PDB file of the protein must be checked to avoid any double occupancies or residue insertions. This can be done manually or using for example the PDB cleaner website (http://www.igs.cnrs-mrs.fr/Caspr2/magicPDB.cgi) [37]. The input files for both the protein and the ensemble of models for the peptides are provided in supplementary material, respectively, named `1NX1_protein.pdb` and `DAIDALSSDFT_3conformations.pdb`.

3.1.2 Docking the Capstatin Peptide onto Capsain with the HADDOCK Web Server

For this docking, we will make use of the Guru interface of the HADDOCK web server (http://haddock.science.uu.nl/services/HADDOCK/haddockserver-guru.html). Note that the Guru interface is available for registered users with appropriate access rights.

1. Open an Internet browser and go to haddock.science.uu.nl/services/HADDOCK. Choose the Guru interface. You will find the page illustrated in Figs. 1 and 2.
2. We advise to give a name to your docking run. Be aware that no space or special characters other than "-" or "_" are allowed. We propose here to name the run 1NX1_modeling.
3. The PDB file of the largest molecule, in this case the calpain domain IV, has to be entered first (*see* **Note 5**). Expand the section `First molecule`. At the entry `Where is the structure provided?` click on the drop-down menu next to it and select `I am submitting it`. Set `Which chain of the structure must be used?` to *A*ll (*see* **Note 4**). Next to `PDB structure to submit` press the `Browse...` button and move to the location where the tutorial data were unpacked. Go to the `pdbs/` directory and select the `1NX1_protein.pdb` file.
4. Specify the interface by defining active and passive residues. We listed the residues that are considered active in Table 1. Fill in the numbers of the active residues in the textbox next to `Active residues`.
5. Specify the `Segment ID to use during the docking` for the first molecule as A (*see* **Note 4**).
6. We leave the proteins flexibility settings to their defaults values: no residues will be considered as fully flexible and semi-flexible segments will be determined automatically by HADDOCK.
7. Both N-terminus and C-terminus of the protein will be considered as charged, the default value (*see* **Note 6**).
8. Expand the `Second molecule` section. The peptide will require some specific settings, which we will explain in the following steps. If a parameter is not mentioned in the following steps, its default value should be kept.
9. At the entry `Where is the structure provided?` click on the dropdown menu next to it and select `I am submitting it`. Set `Which chain of the structure must be used?` to `All` (*see* **Note 1**). Next to `PDB structure to submit` press the `Browse...` button and move to the location where the tutorial data were unpacked. Go to the `pdbs/` directory and select the `DAIDALSSDFT_3conformations.pdb` file.
10. As explained before, the entire peptide will be considered as passive during the docking process. For this, enter each residue number present in the peptide PDB (for one model) separated by a comma in the *Passive residues* textbox as indicated in the Table 1 (*see* **Note 10**).

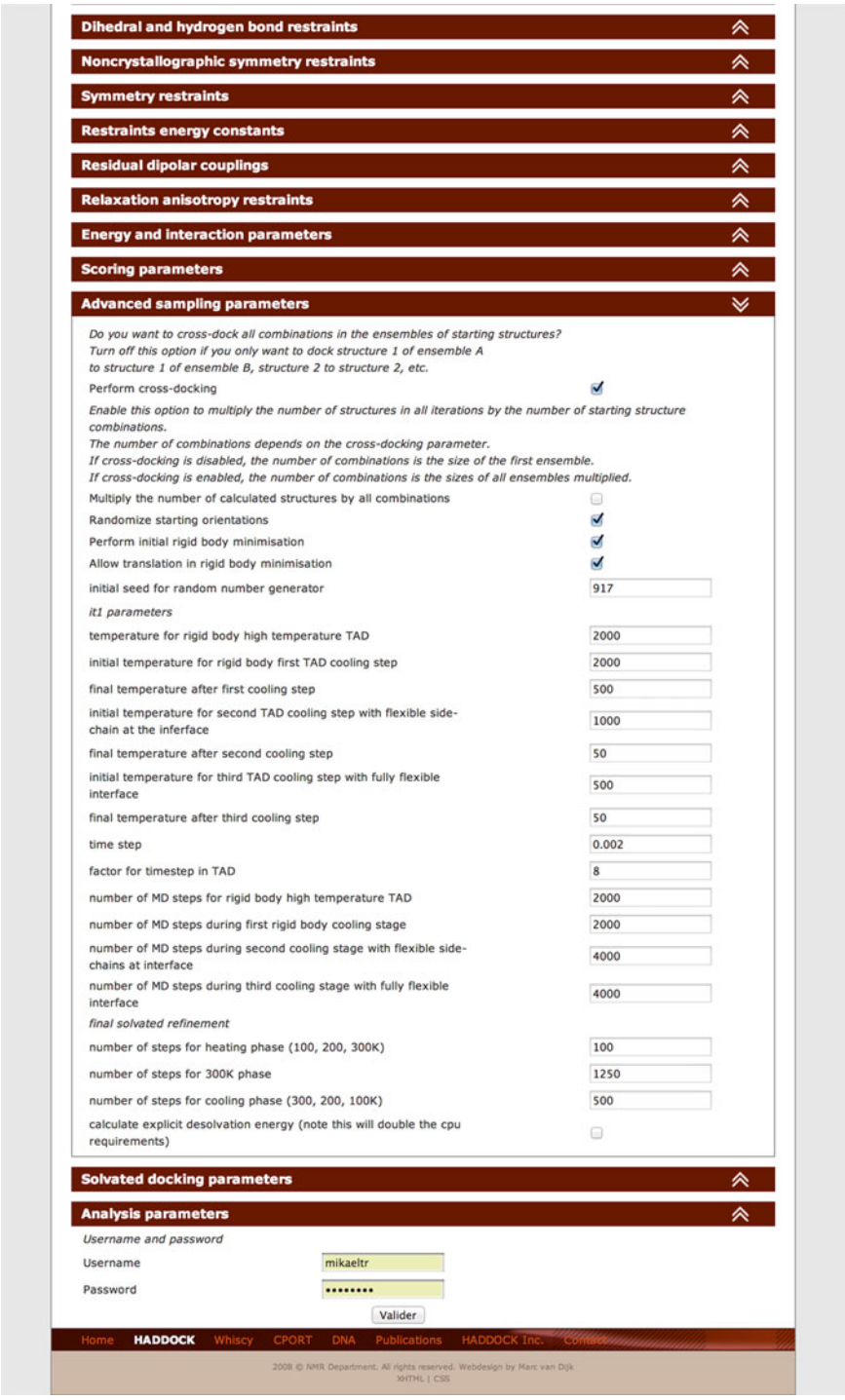


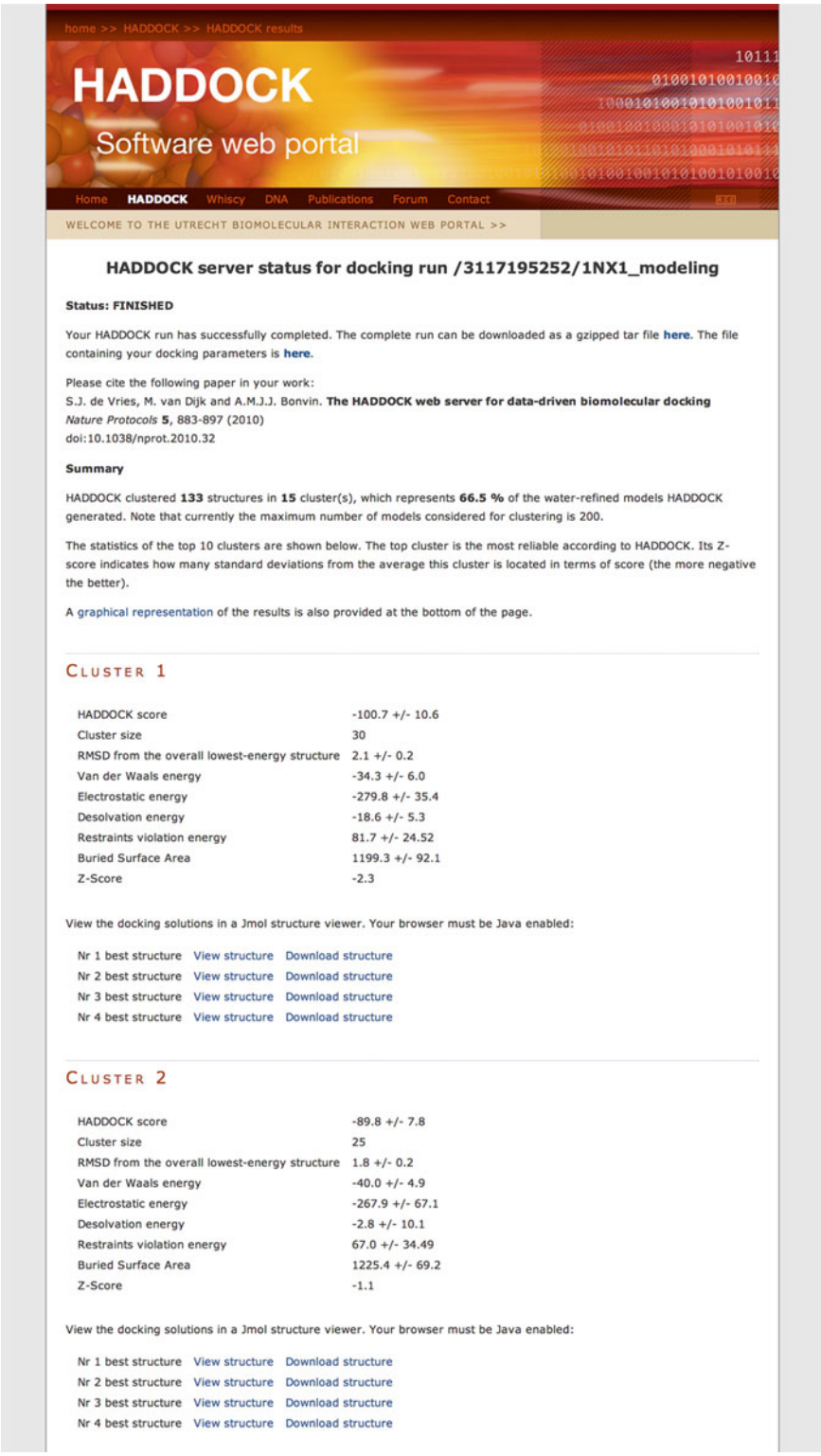
Fig. 2 Overview of the HADDOCK web server Guru interface. The expanded sections include input parameters that need to be changed to perform a protein-peptide docking run. The sections concerned are *Parameters for clustering* and *Advanced sampling parameters*

Table 1
Input data used for the protein-peptide docking run

Protein (Calpain Domain VI)	
Active residues	6, 9, 12, 13, 28, 31, 32, 33, 35, 36, 38, 39, 69, 73, 76, 77, 80, 81, 84, 131
Passive residues	None
Fully flexible segments	None
Peptide (Calpastatin inhibitory domain C)	
Active residues	None
Passive residues	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Fully flexible segments	11-Jan
C- and N-termini	Uncharged

11. Here, the peptide corresponds to a disordered segment of the capstatin protein and should thus be considered as noncharged at the Cter and Nter. For this, uncheck the two boxes, respectively, *The C-terminus of your protein is negatively charged* and *The N-terminus of your protein is negatively charged* (*see* **Note 6**).
12. In our example, no input data other than the list of active residues on the protein receptor will be used.
13. In the *Sampling parameters* section, we increase the *Number of structures for rigid body docking* (it0) from 1,000 to 6,000. In that way, each conformation is sampled 2,000 times in the rigid body stage. We also increase the *Number of structures for semi-flexible refinement* (it1) and the *Number of structures for the explicit solvent refinement* (water) to 400 structures.
14. Go to the *Parameters for clustering* section and change the *RMSD Cutoff for clustering* from 7.5 to 5.0 to make up for the smaller size of protein-peptide interfaces.
15. In the *Advanced sampling parameters* section, the default numbers of MD steps are multiplied by a factor 4 to increase the sampling. Therefore, the *number of MD steps for rigid body high temperature TAD*, the *number of MD steps during first rigid body cooling stage*, the *number of MD steps during second cooling stage with flexible side-chains at interface* and *the number of MD steps during third cooling stage with fully flexible interface* are respectively set to 2,000/2,000/4,000/4,000.

16. You can now fill in your `Username` and `Password` at the bottom of the submission page and click on the `Submit Query` button. After few seconds you will be redirected to a page reporting the status of your job, first the outcome of the validation steps performed by the HADDOCK web server, then a link to the result page and the possibility to download a unique self-contained file to resubmit your job (provided here with the default name `haddockparam.web`). On the result page, you can monitor the progress of your docking run. When finished, it will later display the final results, which consist in generic analyses of the models. An email to confirm the processing of your job is sent to your registration email address.
17. Within typically a couple of hours, depending on the web server load, you will receive another email reporting the final status of your job. If successful, a result page as depicted in Fig. 3 will be available at the link given in the e-mail. On this page, you will find the name of your docking run as well as a link to download it as a gzipped tar file. A link to the unique file containing input data and parameters is again provided.
18. In this page, you will find the number of clusters created by HADDOCK and how many structures coming from the *water* steps have been clustered. By default, only the 200 models with the lowest HADDOCK scores are analysed, therefore only half of the refined models are clustered. In our example, 15 clusters are created, gathering 66.5 % of the top 200 models. For an easier visualization of the results, only the ten best clusters based on the average HADDOCK score of its top four models are displayed in the summary page. You can find information and analyses of the last cluster in the gzipped tar file. For each cluster, information relative to the HADDOCK score of the top four models, the cluster size and different statistics and energy values are reported (*see* **Note 7**).
19. At last, a graphical representation of different CAPRI assessment criteria with respect to the HADDOCK score is provided for the ten best clusters in the `Results analysis` section as shown in Fig. 4. The first three plots show the HADDOCK score versus the interface-ligand-RMSD (i-l-RMSD), the i-RMSD and the l-RMSD, respectively (*see* **Note 8**). The next plot displays the HADDOCK score versus the fraction of common contacts (FCC) (*see* **Note 9**). The last three plots show the van der Waals, electrostatics, and AIRs energy versus i-RMSD.
20. It is possible to manually compare a reference structure with the best models of each cluster generated by HADDOCK. The 3D structures of these models are located in the root of the docking run you downloaded as a gzipped tar file. Their name follows the following syntax: `cluster2_1.pdb`.



home >> HADDOCK >> HADDOCK results

HADDOCK

Software web portal

Home HADDOCK Whiscy DNA Publications Forum Contact

WELCOME TO THE UTRECHT BIOMOLECULAR INTERACTION WEB PORTAL >>

HADDOCK server status for docking run /3117195252/1NX1_modeling

Status: FINISHED

Your HADDOCK run has successfully completed. The complete run can be downloaded as a gzipped tar file **here**. The file containing your docking parameters is **here**.

Please cite the following paper in your work:
S.J. de Vries, M. van Dijk and A.M.J.J. Bonvin. **The HADDOCK web server for data-driven biomolecular docking** *Nature Protocols* **5**, 883-897 (2010)
doi:10.1038/nprot.2010.32

Summary

HADDOCK clustered **133** structures in **15** cluster(s), which represents **66.5 %** of the water-refined models HADDOCK generated. Note that currently the maximum number of models considered for clustering is 200.

The statistics of the top 10 clusters are shown below. The top cluster is the most reliable according to HADDOCK. Its Z-score indicates how many standard deviations from the average this cluster is located in terms of score (the more negative the better).

A graphical representation of the results is also provided at the bottom of the page.

CLUSTER 1

HADDOCK score	-100.7 +/- 10.6
Cluster size	30
RMSD from the overall lowest-energy structure	2.1 +/- 0.2
Van der Waals energy	-34.3 +/- 6.0
Electrostatic energy	-279.8 +/- 35.4
Desolvation energy	-18.6 +/- 5.3
Restraints violation energy	81.7 +/- 24.52
Buried Surface Area	1199.3 +/- 92.1
Z-Score	-2.3

View the docking solutions in a Jmol structure viewer. Your browser must be Java enabled:

Nr 1 best structure View structure Download structure
Nr 2 best structure View structure Download structure
Nr 3 best structure View structure Download structure
Nr 4 best structure View structure Download structure

CLUSTER 2

HADDOCK score	-89.8 +/- 7.8
Cluster size	25
RMSD from the overall lowest-energy structure	1.8 +/- 0.2
Van der Waals energy	-40.0 +/- 4.9
Electrostatic energy	-267.9 +/- 67.1
Desolvation energy	-2.8 +/- 10.1
Restraints violation energy	67.0 +/- 34.49
Buried Surface Area	1225.4 +/- 69.2
Z-Score	-1.1

View the docking solutions in a Jmol structure viewer. Your browser must be Java enabled:

Nr 1 best structure View structure Download structure
Nr 2 best structure View structure Download structure
Nr 3 best structure View structure Download structure
Nr 4 best structure View structure Download structure

Fig. 3 Example view of a result page of the HADDOCK web server. Links toward the complete run and a HADDOCK-formatted summary of your input parameters can be found. Moreover, a brief summary of the clustering performances is shown with a focus on the first two clusters (according to HADDOCK score average of the top-four structures) analytical information

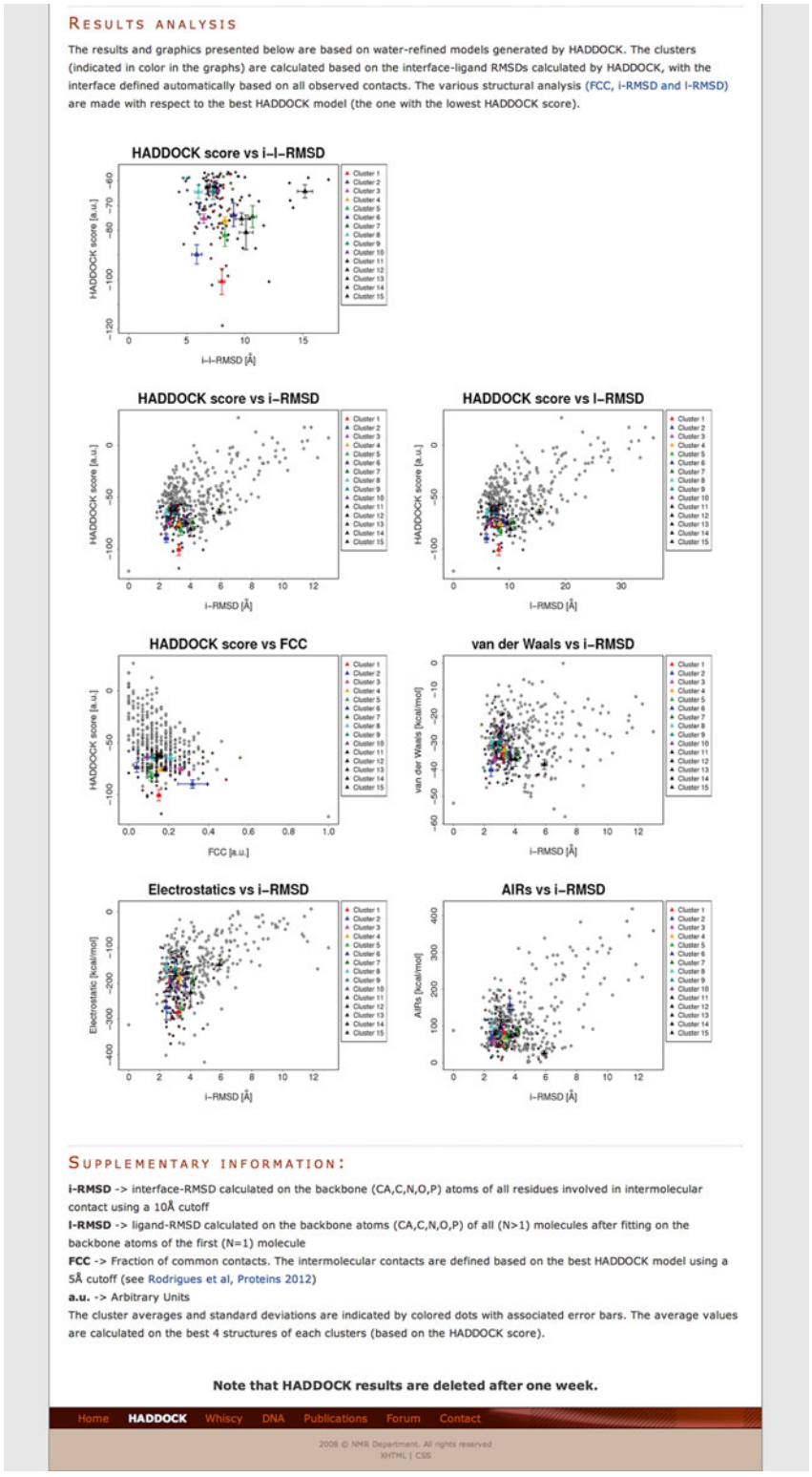


Fig. 4 Results analysis section of a result page of the HADDOCK web server. Several graphics with the main energetic parameters plotted with respect to the HADDOCK score are shown and separated according to the cluster number of each structure

This file is for instance the best model according to its HADDOCK score in the second cluster given by HADDOCK.

You can use ProFit to get precise values of RMSD. PyMol is useful as well since it has its own fitting algorithm and will give you a RMSD value as well as a visual feedback of the differences between the clustered models and the reference structure. Keep in mind that your reference structure has to be formatted in the same way that the PDB models generated by HADDOCK. ProFit considers only structures with an identical number of atoms.

4 Case Studies

The settings we described before have been used to test HADDOCK against a large benchmark of 62 protein-peptide complexes for which an unbound form of the protein was available. The challenge was then double here: model successfully the peptide's conformation at the correct interface and reproduce the bound form of the protein. We analysed the quality of the models generated by HADDOCK but also our capacity to rank efficiently the correct predictions among the top HADDOCK score models. HADDOCK successfully generated acceptable models (*see* Subheading 2.3.3 for definition of acceptable models) for about 70 % of the tested cases (Fig. 5a). Among these (*see* **Note 9**), at least one acceptable model or better is found in the top 20 models in 76 % of the cases. But after clustering, 50 % of the cases contain at least an acceptable structure in the best cluster and this quickly reaches 75 % if the top three clusters are considered (Fig. 5b).

We illustrated the HADDOCK protein-peptide docking protocol in Subheading 3.1.2 with the modeling of the calpain/calpastatin complex, starting from the unbound structure of the calpain and an ensemble of three conformations for the disordered region of the calpastatin. In the last step of HADDOCK protocol (refinement water step), this docking run generated 25 final acceptable models (i-RMSD ≤ 2 Å), 18 of which ended in a cluster and 7 were not clustered. Among the 18 clustered models, 15 come from the 2nd best cluster according to HADDOCK and two structures are the 1st and 4th models based on their HADDOCK score. The 2nd best cluster given by HADDOCK is the cluster for which the average HADDOCK score of its top four models is the second lowest among all the clusters. To get a precise idea, the best cluster has an average HADDOCK score for its four best models of -100.7 ± 10.6, as opposed to -89.8 ± 7.8 for the 2nd best cluster. Considering the standard deviations those tow clusters are rather close. The representative best four models of the top-ranking cluster have, on average, an i-RMSD of 3.9 ± 0.3 Å when compared to the crystal structure whereas the best four models of the second best cluster have an average i-RMSD of 1.9 ± 0.3 Å.

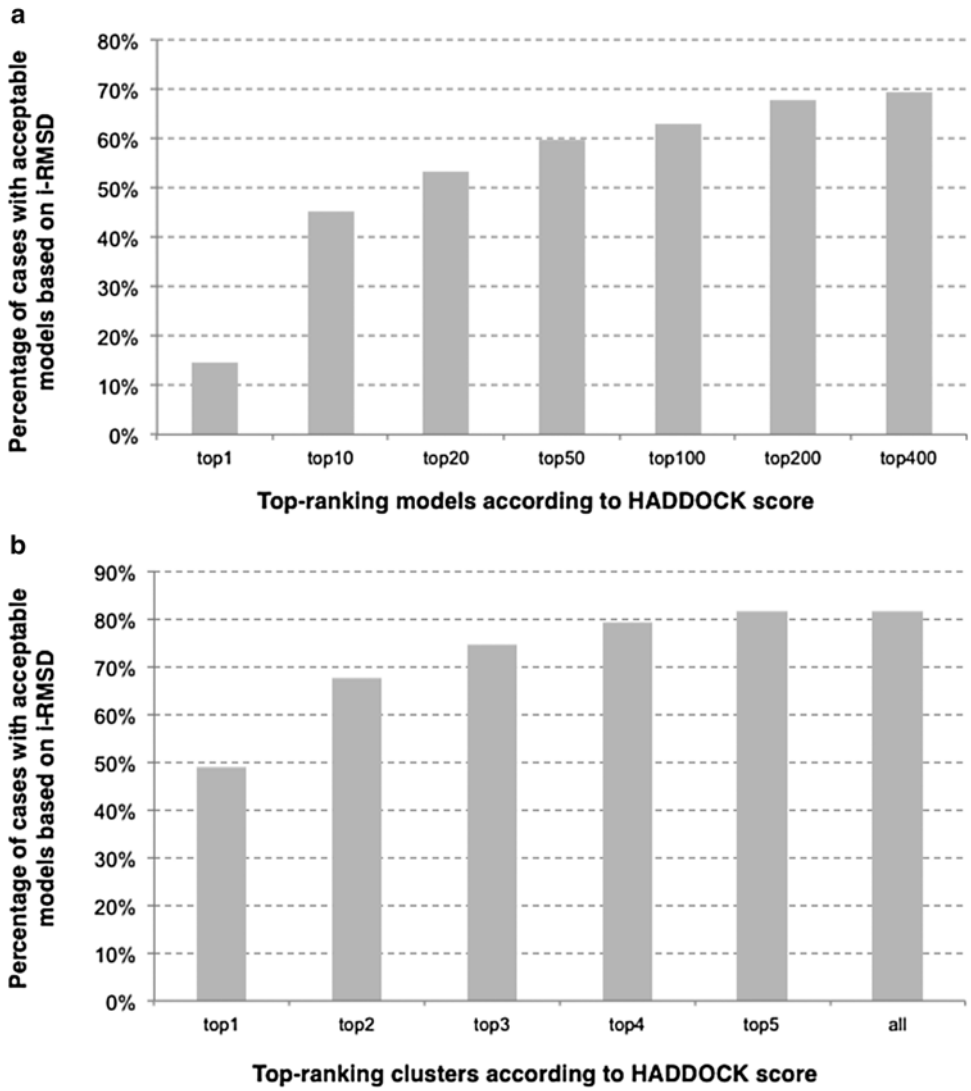


Fig. 5 (**a**) Success rate of unbound/unbound docking as a function of the number of top models considered. (**b**) Clustering performance of HADDOCK in unbound/unbound docking onto acceptable cases (with at least one acceptable model) as a function of the number of clusters considered

The peptide starting conformations and the resulting best model in term of i-RMSD from the reference complex are shown in Fig. 6. Statistics of the two clusters is presented in Table 2. We voluntarily chose this case to illustrate that the correct solution is not always on top and various clusters should be examined, especially when their scores are rather close. Ideally, it would be best to have some independent data at hand to validate the generated models. The models can also serve at starting point for the design of experiments to test the predictions, for example by mutagenesis. It is often the synergistic combination of modeling and experiment that allows to answer challenging biological questions.

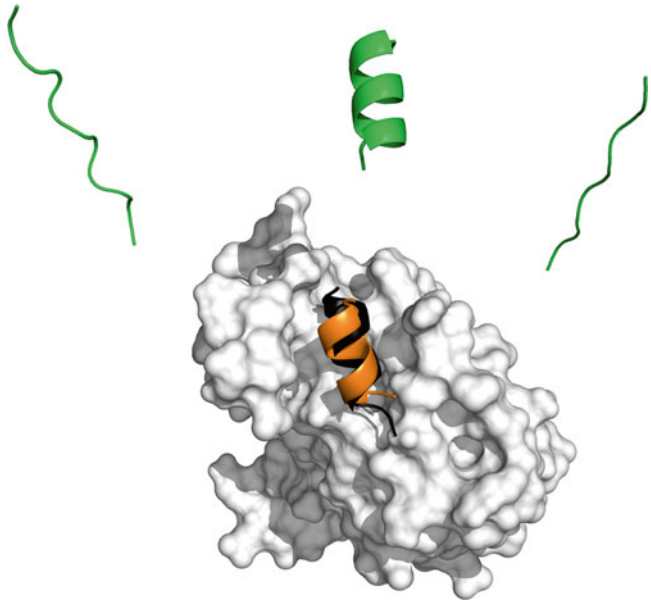


Fig. 6 Summary of HADDOCK protocol illustrated by the docking run between the calpain domain VI and the calpastatin inhibitory domain C. In *green*, the three starting conformations provided to HADDOCK. In *white*, the protein (calpain) rendered as a surface, in *black* the crystal conformation of the peptide as found in the bound complex PDB file and in *purple*, a near-native model corresponding to HADDOCK's third best ranked structure (here the first model of the second cluster)

Table 2
Comparison of two best clusters from HADDOCK for 1NX1 modelling run

	Cluster 1	Cluster 2
HADDOCK score (average)	−100.7 ± 10.6	−89.8 ± 7.8
Cluster size	30	28
RMSD from the overall lowest-energy structure (Å)	2.1 ± 0.2	1.8 ± 0.2
Z-score	−2.3	−1.1
i-RMSD from reference structure for best four structures (Å)	3.9 ± 0.3	1.9 ± 0.3

5 Notes

1. Note that the server does support N-acetylated and C-amino termini. This can however not be specified in the web form, but must be done by editing the coordinates file of the peptide molecule and adding residues at the N- and C-termini, respectively, named ACE/CTN (for example adding a GLY at the termini and rename it to ACE/CTN, respectively; HADDOCK will take care of removing/adding the necessary atoms).

2. This feature will be available in the next release of the web server and is available in the local installation of HADDOCK upon request.
3. For very short peptides, this value might be further decreased.
4. The PDB files provided to HADDOCK have to be correctly formatted to avoid any issues during the simulation process. Any chainID and/or segID should be removed from the input PDBs and there should be no overlap in residue numbering. This can be done for example using the PDB cleaner website (http://www.igs.cnrs-mrs.fr/Caspr2/magicPDB.cgi) [37]. Missing atoms in the PDB files are not problematic since HADDOCK will rebuild them based on the topology files of the force field.
5. Defining the largest molecule as first molecule for docking is important for the final clustering because the structures are first fitted on the interface residues of the first molecule and then the RMSD is calculated on the interface residues of the second molecule. The interface residues are defined from an analysis of contacts in the generated models (at it1 and water, respectively). Defining the largest molecule first should thus result in a better fitting.
6. The charge state of the termini has to be properly set depending on the system under study: naturally occurring peptide with charged termini or peptide fragment extracted from a protein (typically loop or intrinsically disordered region), which should be uncharged … (*see* also **Note 1**).
7. The Z-score indicates how many standard deviations from the average a cluster is located in terms of its HADDOCK score. So the more negative the better.
8. All reported RMSDs are calculated with respect to the lowest scoring model (the best model according to the HADDOCK score). The i-l-RMSD, which is used for clustering, is calculated on the interface backbone atoms of all chains except the first one after fitting on the backbone atom of the interface of the first molecule. The i-RMSD is calculated by fitting on the backbone atoms of all the residues involved in intermolecular contacts within a cutoff of 10 Å. The l-RMSD is obtained by first fitting on the backbone atoms of the first molecule and then calculating the RMSD on the backbone atoms of the remaining chains.
9. The FCC stands for Fraction of Common Contacts and is calculated by comparing the lists of contacts at the interface between the protein and the peptide chain in the reference structure and the model structure. A contact is defined when two residues from different chains of the complex are closer than 5 Å from each other. The FCC is then the percentage of common residue pairs shared between a model and the reference structure.
10. We define a successful case a case for which at least one acceptable model is present in the final 400 models generated.

References

1. Tzakos AG, Fuchs P, van Nuland NA et al (2004) NMR and molecular dynamics studies of an autoimmune myelin basic protein peptide and its antagonist: structural implications for the MHC II (I-Au)-peptide complex from docking calculations. Eur J Biochem 271: 3399–3413
2. Musi V, Birdsall B, Fernandez-Ballester G et al (2006) New approaches to high-throughput structure characterization of SH3 complexes: the example of Myosin-3 and Myosin-5 SH3 domains from S. cerevisiae. Protein Sci 15: 795–807
3. Huang BX, Kim H-Y (2006) Interdomain conformational changes in Akt activation revealed by chemical cross-linking and tandem mass spectrometry. Mol Cell Proteomics 5:1045–1053
4. Casares S, Ab E, Eshuis H et al (2007) The high-resolution NMR structure of the R21A Spc-SH3:P41 complex: understanding the determinants of binding affinity by comparison with Abl-SH3. BMC Struct Biol 7:22
5. Gelis I, Bonvin AM, Keramisanou D et al (2007) Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. Cell 131:756–769
6. Schneider T, Kruse T, Wimmer R et al (2010) Plectasin, a fungal defensin, targets the bacterial cell wall precursor Lipid II. Science 328: 1168–1172
7. Wodak SJ, Janin J (1978) Computer analysis of protein-protein interaction. J Mol Biol 124:323–342
8. Strynadka NCJ, Eisenstein M, Katchalski-Katzir E et al (1996) Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. Nat Struct Mol Biol 3:233–239
9. Petsalaki E, Russell RB (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 19:344–350
10. Stein A, Aloy P (2008) Contextual specificity in peptide-mediated protein interactions. PLoS One 3:e2524
11. London N, Movshovitz-Attias D, Schueler-Furman O (2010) The structural basis of peptide-protein binding strategies. Structure 18:188–199
12. London N, Raveh B, Schueler-Furman O (2013) Peptide docking and structure-based characterization of peptide binding: from knowledge to know-how. Curr Opin Struct Biol 23:894–902
13. Petsalaki E, Stark A, Garcia-Urdiales E, Russell RB (2009) Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput Biol 5:e1000335
14. Antes I (2010) DynaDock: a new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. Proteins 78:1084–1104
15. Raveh B, London N, Schueler-Furman O (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins 78:2029–2040
16. Ben-Shimon A, Eisenstein M (2010) Computational mapping of anchoring spots on protein surfaces. J Mol Biol 402:259–277
17. Dagliyan O, Proctor EA, D'Auria KM et al (2011) Structural and dynamic determinants of protein-peptide recognition. Structure 19: 1837–1845
18. Raveh B, London N, Zimmerman L, Schueler-Furman O (2011) Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. PLoS One 6:e18934
19. Donsky E, Wolfson HJ (2011) PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors. Bioinformatics 27: 2836–2842
20. Lavi A, Ngan CH, Movshovitz-Attias D et al (2013) Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions. Proteins 81:2096–2105
21. Verschueren E, Vanhee P, Rousseau F et al (2013) Protein-peptide complex prediction through fragment interaction patterns. Structure 21:789–797
22. De Vries SJ, van Dijk AD, Krzeminski M et al (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins 69:726–733
23. Trellet M, Melquiond ASJ, Bonvin AMJJ (2013) A unified conformational selection and induced fit approach to protein-peptide docking. PLoS One 8:e58769
24. Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737
25. Brünger AT, Adams PD, Clore GM et al (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921
26. Brunger AT (2007) Version 1.2 of the crystallography and NMR system. Nat Protoc 2: 2728–2733
27. Jorgensen WL, Tirado-Rives J (1988) The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 110:1657–1666

28. Moreira IS, Fernandes PA, Ramos MJ (2010) Protein-protein docking dealing with the unknown. J Comput Chem 31:317–342
29. Lensink MF, Wodak SJ (2013) Docking, scoring, and affinity prediction in CAPRI. Proteins 81:2082–2095
30. Diella F, Haslam N, Chica C et al (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13:6580–6603
31. Van Dijk ADJ, Boelens R, Bonvin AMJJ (2005) Data-driven docking for the study of biomolecular complexes. FEBS J 272:293–312
32. Melquiond ASJ, Bonvin AMJJ (2010) Data-driven docking: using external information to spark the biomolecular rendez-vous. In: Protein-protein complexes: analysis, modelling and drug design. Edited by M. Zacharrias, Imperial College Press, London, p 183–209
33. Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
34. Janin J, Henrick K, Moult J et al (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52:2–9
35. Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073–3084
36. Schrodinger L (2010) The PyMOL molecular graphics system, version 1.3r1
37. Claude J-B, Suhre K, Notredame C et al (2004) CaspR: a web server for automated molecular replacement using homology modelling. Nucleic Acids Res 32:W606–W609