# Dissecting Biomolecular Interactions by Integrative Modeling

Ezgi Karaca

ISBN/EAN 978-90-5335-644-9

Doctoral Thesis
Dissecting Biomolecular Interactions by Integrative Modeling
Ezgi Karaca
Computational Structural Biology Group, NMR Spectroscopy Research Group,
Bijvoet Center for Biomolecular Research, Faculty of Science / Chemistry,
Utrecht University, The Netherlands
February 2013


Printed in the Netherlands by Ridderprint BV

# Dissecting Biomolecular Interactions by Integrative Modeling

# Analyse van biomoleculaire interacties doormiddel van geïntegreerd modelleren

(met een samenvatting in het Nederlands)

**Proefschrift**

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 6 februari 2013 des ochtends te 10.30 uur

door

**Ezgi Karaca**

geboren op 29 juli 1983
te Istanboel, Turkije

**Promotor:**

Prof. Dr. Alexandre M.J.J. Bonvin

**Beoordelingscommissie:**

Prof. Dr. Marc Baldus

Prof. Dr. Ineke Braakman

Prof. Dr. Türkan Haliloğlu

Prof. Dr. Albert J.R. Heck

*dedicated to my beloved family*
*who always encouraged me*
*to follow my dreams*

## Table of Contents

## List of Abbreviations

| | |
|---|---|
| 3D | Three-Dimensional |
| AFM | Atomic Force Microscopy |
| CCS | Collision Cross Section |
| EM | Electron Microscopy |
| EPR | Electron Paramagnetic Resonance |
| FRET | Förster resonance energy transfer |
| IM | Ion Mobility |
| RMSD | Root Mean Square Deviation |
| MD | Molecular Dynamics |
| MS | Mass Spectrometry |
| NMR | Nuclear Magnetic Resonance |
| NOE | Nuclear Overhauser Effect |
| PCS | Pseudocontact Shifts |
| PRE | Paramagnetic Relaxation Enhancement |
| SA | Simulated Annealing |
| STEM | Scanning Transmission Electron Microscopy |

**“One thing I have learned in a long life: All our science, measured against reality, is primitive and childlike — and yet it is the most precious thing we have.”**

*Albert Einstein*

Chapter

1

# General Introduction: Integrative Modeling of Biomolecular Complexes

Based on the research article:

**Ezgi Karaca, Alexandre M.J.J. Bonvin,** Advances in Integrative Modeling of Biolmolecular Complexes, *Methods*, 2012, *in press*.

## 1. Integrative Modeling of Biomolecular Complexes

Proteins and their intricate network of interactions are the mainstay of any cellular process. Dissecting their interaction networks at atomic detail is therefore invaluable, as this will pave the route to a mechanistic understanding of biological function. Atomic detail (high-resolution) information about structure and dynamics of biomolecular complexes is typically acquired by classical experimental methods such as X-Ray crystallography and NMR spectroscopy. Compared to other structural biology methods, these are the most accurate ones. They are, however, faced with many challenges, especially when the macromolecular systems under study become very large, comprise flexible or unstructured regions, exist in very tiny amounts, are membrane associated, or when their constituents interact only transiently. In the last decade another method, cryo-EM has emerged as a promising alternative for (high-resolution) structure determination. Its advantage over classical techniques is that it does not require high sample concentration [1,2], leading routinely to medium resolutions in the 8-20Å range [3]. But rarely the resolution gets better than 8 to 6Å, which could only be obtained so far for highly symmetric and stable complexes [4–6].

The number of known 3D structures of macromolecular complexes is considerably smaller than the amount of documented protein-protein interaction data [7,8]. Technical limitations of high-resolution methods and other problems mentioned above hamper closing this growing gap in a rapid manner. As a rescue strategy, structural biologists often resort to using different types of biochemical and biophysical experiments that can quickly provide accurate low-resolution information even for challenging systems. Most of the time, however, the collected data are rather sparse and/or of limited information content. These limitations call for integrative computational tools, like for example docking, that can, using some kind of physical model, judiciously combine and accurately translate sparse experimental data into structural information [9–11].

Currently, integrative modeling is the best strategy when conventional structural methods fail. Using such an integrative approach should reduce the downside features of both experimentation and modeling. From an experimentalist point of view, integrative modeling is beneficial since it can generate new hypothesis to drive experiments, which can significantly speed up the structure determination process and/or increase our understanding of biological function [10–12]. It is also advantageous for modelers, as incorporating experimental data into the modeling can accelerate the computational search and greatly help to overcome the shortcomings of *ab initio* modeling, such as high rates of false positives and difficulties in assessing the accuracy of the generated models [13,14].

Integrative methods have most of the time been developed with the drive of dissecting a specific system. Recent examples include successful characterization of a wide range of challenging systems, varying from flexible dimers to whole cells, based on different combinations of X-Ray, NMR, cryo-EM, Electron Tomography and SAXS data [15–18]. All of these are important milestones in the field of integrative modeling, however, being mainly application-oriented or system-specific, their general applicability still has to be demonstrated [17]. There is a small number of generic integrative modeling approaches and these are the main focus of this review. We discuss them in detail in the following sections. In the final section, we concentrate on our data-driven docking approach, HADDOCK, and position it within the current state of generic integrative modeling methods by presenting some application examples.

## 2. Translating sparse data into 3D structures

### *2.1. Sources of low-resolution information*

There are various types of biophysical and biochemical experimental techniques that can quickly provide low-resolution structural information. Assuming

that the stoichiometry and composition of the macromolecular complex is known, these can provide useful insight into binding sites, distances between specific pair or groups of atoms, orientation between molecules and/or globular shape of the complex. The most frequently used data and their information content are summarized in **Table 1**.

Chemical Shift Perturbation (CSP), Hydrogen/Deuterium (H/D) exchange, solvent Paramagnetic Relaxation Enhancement (PRE) and chemical footprinting experiments provide information about interacting surfaces [11,12,17,19]. They all determine the binding site based on the alteration of the environment upon complexation. CSP measures the chemical environment changes induced by ligand binding [20–22]. H/D exchange is conducted by monitoring the exchange of labile hydrogens with deuteriums, so that changes in surface accessibility can be detected [23,24]. Solvent PREs are measured by using chemically inert paramagnetic probes as co-solvents that cause relaxation and thus signal attenuation of solvent accessible protons [25]. In chemical footprinting, the non-interacting surface of the complex is exposed to chemical modification, leaving the binding site unaffected [26]. Mutagenesis allows to identify specific residues that are critical for binding [12,27]. Next to those methods, bioinformatics approaches based on sequence/structure conservation [28], comparative patch analysis [29], correlated mutation studies [30], possibly combined with information about surface properties (e.g. curvature, hydrophobicity, charge) [31], can also be used to predict binding sites. All these approaches are built on the idea that conservation of sequence, contacts, patches or a globular structural element can possibly depict a probable interaction site [11,32,33].

Short-range distances between pair of atoms can be obtained by NOE measurements (<5-6 Å) [34,35], which are used together with dihedral angle restraints derived from J-couplings measurements or from chemical shifts analysis in conventional NMR structure calculation [36]. Chemical cross-linking experiments provide another source of distance information [37,38]. In these, functional groups on the surface of biomolecules are cross-linked using reactive chemicals. Residues are cross-linkable, if they are in close proximity and have chemical properties (e.g. Lys side-chains) allowing them to bind covalently to the cross-linking agent. They are usually identified by MS. The measured distance ranges depend on the cross-linker size and flexibility [39,40]. In the presence of paramagnetic ions (e.g. substituted in a metal binding site or attached to the protein via a tag), PRE [15], Pseudocontact Shifts (PCS) NMR [41] measurements or EPR experiments [42] can help to identify long-range distances up to 20 to 40 Å, depending in the paramagnetic species used and even 80 Å for EPR measurements. PCS, in addition, also contain orientational information. These effects are observed due to magnetic dipolar interactions between

nucleus and the unpaired electrons of the paramagnetic center [17,43,44]. FRET experiments provide another source of long-range distance information: the measured distances depend on the separation of the fluorescently labeled residues of the complex [45–47].

**Table 1. Sources of low-resolution data, classified based on their information content.**

| | Data type | Experimental technique |
|---|---|---|
| **Binding Site** | Chemical shift perturbations[a] | NMR |
| | H/D exchange[a] | NMR, MS |
| | Solvent PRE[a] | NMR |
| | Mutagenesis[a] | Biochemistry |
| | Chemical footprinting[a] | Biochemistry |
| | Conservation (Correlated mutations, Comparative patch analysis)[a] | Bioinformatics predictions |
| **Distance** | NOE distances[b] | NMR |
| | Chemical cross-links[b] | MS |
| | PRE[b] | NMR |
| | Correlated mutations[c] | Biochemistry |
| | PCS[b] | NMR |
| | Distances (distribution)[d] | FRET |
| | Distances from EPR[d] | EPR |
| **Orientation** | Residual dipolar couplings[e] | NMR |
| | Relaxation anisotropy[e] | NMR |
| **Shape** | Collision cross section[f] | IM-MS |
| | Radius of gyration[f] | SAXS |
| | Molecular envelope, globular shape[f] | SAXS, cryo-EM |

Resolution/ambiguity level for a given source of information:
[a]Residue; [b]Atom-atom (separation); [c]Residue-residue (separation); [d]Label-label (separation); [e]Inter-monomer and/or bond vector orientations; [f]Biomolecular complex.

Information on the relative orientation of two molecules can be obtained by Residual Dipolar Coupling (RDC) [48] or NMR Relaxation experiments [49]. In conventional NMR structure calculations, this orientational information is often combined with binding site and distance information, from CSP's and NOE's,

respectively [19,35]. Lately it has also been frequently used with shape data from SAXS experiments, in order to reduce the inherent degeneracy entailed by the low-resolution shapes [50,51]. Low-resolution shape information can be obtained from SAXS and cryo-EM experiments. SAXS experiments measure the scattering intensity at very low angles, which can be translated into a low-resolution 3D envelope [52]. In addition, the radius of gyration (Rg) of a complex can be extracted from the SAXS data, which is an indicator of the structure compactness [53]. Cryo-EM experiments provide an electron density map with a resolution range typically between 8 and 20 Å [1,3]. The molecular maps extracted from cryo-EM experiments are especially useful when the individual structures of a complex's constituents are known, since these can then be fitted into those maps [54]. Finally, IM-MS experiments also provide shape-related information in the form of Collision Cross Sections (CCS). The CCS corresponds to the rotationally–averaged molecular area to which the buffer gas can collide; it offers thus information on the overall size and conformation of the complex [55–57].

For further information and a more detailed description of the various types of experimental data, please refer to the comprehensive review of *Melquiond* and *Bonvin* on data-driven modeling [12].

### *2.2. Integration of sparse data into modeling*

Integrative modeling of complexes is typically composed of two stages. The first stage is sampling, where the conformational space is searched and the second one is scoring, in which the generated models are ranked based on some energy function. Sampling can be accomplished by minimizing a target energy function or by searching for geometric surface correlations [58–61]. The latter ensures an exhaustive sampling of the rotational and translational space, in order to find the binding mode that provides the best surface complementarity. Common surface correlation methods are based on Fast Fourier Transformations [62,63] or geometric hashing [64]. Minimization/optimization methods are often not exhaustive; they typically perform a nonlinear optimization of a defined target function, which encodes biophysical/biochemical properties of the biomolecules and their interfaces [65–67]. Gradient-based energy minimization [68] or molecular dynamics methods [69], or Metropolis Monte Carlo simulations [70], which only require energy calculations, are the most frequently used optimization methods [59,60,71]. There are a number of generic modeling approaches making use of these sampling methods in a versatile manner. One of them is docking, where the conformational search is particularly dedicated to identifying the correct macromolecular interface and reproducing its physicochemical properties.

The experimental and/or bioinformatics data can be integrated into macromolecular modeling either *a priori* during sampling, by restraining the conformational search space, or *a posteriori* during scoring, by filtering or ranking the generated models based on their fit to the experimental data [58,72]. In the latter case, the conformational search should be done globally by thorough exploration of the interaction space (**Figure 1A**). This type of search typically results in a large number of heterogeneous models. However, if information is used to drive the conformational search (i.e. *a priori*), the search can be concentrated on a fraction of the interaction space, defined by the input data, thus resulting in an often more homogenous set of solutions (**Figure 1B**).

A straightforward way of imposing a restraint *a priori* is to incorporate it as an additional energy term into the existing force field (Eq. 1).

$$E_{target} = E_{FF} + w_{data}E_{restr} \quad \text{(Eq. 1)}$$

The combination of force field ($E_{FF}$) and restraint energy ($E_{restr}$) terms defines the target function ($E_{target}$) that reaches its minimum, when the computed model simultaneously agrees with *a priori* encoded chemical and physical information and the observed data [10,11,71,73]. $E_{FF}$ denotes the empirical knowledge on covalent (e.g. bonds, angles, dihedrals and chirality) and non-bonded (electrostatics and van der Waals) interactions, expressed in molecular mechanics terms [66,74,75]. $E_{restr}$ on the other hand, describes the discrepancy between observed and calculated data. It is often expressed as a harmonic potential [71]:

$$E_{restr} = (R^{Calc} - R^{Refe})^2 \quad \text{(Eq. 2)}$$

where $R^{Calc}$ is the back-calculated value from the structure and $R^{Refe}$ is the experimental reference value. The target value of the quadratic potential can be either a single point or an interval, in between which $E_{restr}$ is 0 [76,77]. This is defined by the nature and precision of the available information. For example, an inter-atomic distance obtained from cross-linking experiments is typically enforced as an interval to account for the flexibility of the linker [39,40]. Each restraining function should be associated with the existing energy terms by choosing a proper weight ($w_{data}$, see Eq. 1), in order to balance the impact of the various energetic terms [71,78]. Further, $E_{restr}$ can be modified to avoid large energies and forces at large violations, which could cause problems in the optimization procedure. Typically the potential function is modified such as to switch smoothly from a harmonic to a linear form after a defined violation. Such functions are often used in NMR structure determination [76,79].

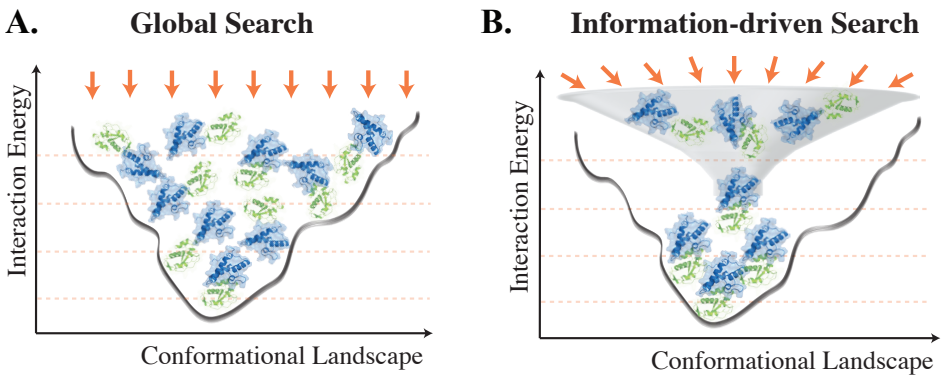



**Figure 1. Global search vs. Information-driven search. A.** A global search method performs a thorough exploration of the conformational space resulting in a large number of heterogeneous solutions. **B.** An information-driven search method is directed by the data supplied and thus the search is only concentrated on limited part of the conformational space. Such a protocol generates a more homogenous set of solutions compared to global search.

After the conformational search, experimental and/or bioinformatics data can be translated into a "fit" term, which measures the discrepancy between the structural properties back-calculated from the generated models and the experimentally measured ones. This term can be used in isolation as an individual filter, so that non-fitting models will be eliminated. It is rather advisable to integrate it with other physicochemical information (e.g. non-bonded energies) into a scoring function [58]. Conventional scoring functions often consist of a weighted sum of based desolvation and contact potential terms [14,58]. The generated models are ranked based on the value of this scoring function. Alternatively, the models can be clustered and then the scores are calculated on a per-cluster basis [80,81]. Individual ranking provides a list of good scoring solutions, whereas, in cluster ranking the solutions are grouped based on a defined similarity measure and ranked according to their average cluster score.

## 3. Challenges of integrative approaches

Albeit the various macromolecular modeling methods differ in their approaches, all of them are confronted with similar challenges due to the intricate conformational space to be sampled, the difficulty of accurate scoring and the ambiguous and degenerate nature of the input data. In the following, we address each of these challenges individually.

### *3.1. What are the challenges?*

#### *3.1.1. Modeling large and dynamic molecular machines*

Large and dynamic molecular machines, such as the proteasome [82], the ribosome [83], the nuclear pore complex and the spliceosome, carry out a majority of essential cellular functions [11,84]. Modeling such assemblies requires, first, being able to deal with multiple molecules, of possibly different natures, and second, being able to cope with conformational changes. These requirements result in a paramount increase in the degrees of freedom and, accordingly, in a highly intricate conformational space to be sampled [85,86].

One of the ways to reduce the number of degrees of freedom is by using coarse-grained representations, in which groups of atoms (or even residues) are represented by a single particle [87–89]. Moreover, if present in the system, inclusion of symmetry considerations will restrain the number of possible conformational poses [33,90,91]. So, in order to efficiently model large macromolecular machines, an ideal integrative approach should be able to handle various types of cyclic and dihedral symmetries and, when needed, should include different levels of coarse-grained representations.

The challenges of dealing with flexibility and conformational changes have been discussed thoroughly in the docking literature, [92–94] revealing that current docking techniques perform well if binding-induced interfacial backbone changes are rather small (≤2 Å). These can be modeled either via refinement of the interface and/or by starting the docking from a suitable ensemble of conformations [67,95–97]. In order to model larger scale conformational changes, incorporation of some experimental data is of great help to simplify the conformational search. For very large changes, however, or even for folding-upon-binding events that, in most cases, will not be sufficient. New methodologies should thus be incorporated to surmount the barriers of exploring the jagged conformational space of the biomolecular interaction [7,98,99].

#### *3.1.2. Constructing an accurate scoring function*

After having generated models/poses by sampling the interaction space, scoring, i.e. fishing out the biologically relevant solutions among the generated pool of conformations is crucial [81]. This is not an easy task. Conventional scoring functions typically describe the thermodynamics and physicochemical properties of the interface through a combination of terms described shortly in **Section 2.2**. Recent analyses have, however, revealed that these functions could not accurately address those properties since they do not correlate with binding free energy [8,14]. This can

be explained in part by the facts that they lack entropic terms and do not take into account the energetics of free components [14,60]. Further, no single scoring function consistently ranks correct solutions at the top [8,14,81,100]. One of the ways to overcome this problem is to incorporate information-based terms into the scoring function [8,14]. Here the choice of the appropriate weight ($w_{data}$, Eq. 1) is critical: too large weights might dominate the scoring and result in unphysical solutions being selected, while too small weights might significantly reduce the effect of the data [71]. Therefore an optimization procedure should be carried out for defining the ideal $w_{data}$ [7]. For example, it has been recently demonstrated that careful incorporation of a SAXS-based term into conventional scoring functions can increase the scoring accuracy significantly [101,102] (see also **Section 4.1**).

*3.1.3. Dealing with the Degeneracy and Ambiguity of the input data*

The main advantage of information-driven approaches over *ab initio* modeling is that the supplied information drives the minimization towards the relevant part of the conformational space, increasing both accuracy and precision. Of course this only holds if the data used are guiding the search towards the relevant part of the conformational space and contain sufficient information, e.g. are non-degenerate, meaning that each member of the set describes a distinct property of the system. Degeneracy, most of the times, arises from an uneven spatial distribution of the data and can directly affect the quality of solutions [11,39,40]. Moreover, the data used might be ambiguous and/or involve false-positives, like in the case of putative binding site information extracted from CSP, H/D exchange and mutation experiments, or even more when bioinformatics predictions are used (see **Table 1)** [103]. Therefore an efficient information-driven method should have a robust energy minimization protocol that can translate non-specific ambiguous data into specific biomolecular interactions, and, it should assess the content of the input data judiciously, by picking, if possible, the relevant (true positive) subset that can drive the structure into a lower energy state or by discarding in some way unreliable, consistently violated data. In order to avoid problems arising from degeneracy and ambiguity, the generated models should be cross-validated against data that have not been used directly during the modeling process. Also, ideally, information from different types of sources should be used in order to increase the data coverage.

### *3.2. Examples of current integrative modeling approaches*

Current integrative methods are dealing with the mentioned challenges in various ways. A short description of a few noteworthy examples is provided below.

Following the technological advancements in the cryo-EM field, numerous EM maps of biomolecular systems have been published in the past decade. Some of these are available from the EM database at the European Bioinformatics Institute (http://www.ebi.ac.uk/pdbe/emdb/ [3]). Consequently, various methods have been developed to fit 3D monomeric structures into those maps [1]. The majority of those focus only on the geometric compatibility between the (low-resolution) EM map and the predicted macromolecular assembly. Multifit [104] distinguishes itself from other the molecules into the EM map. Starting from the bound conformations of the monomers, Multifit could successfully produce near-native models (global-RMSD < 5 Å) for a benchmark of 10 symmetric and asymmetric multi-protein assemblies. The next challenge for Multifit will be to develop a flexible docking protocol to address conformational changes taking place upon binding.

A prominent way to account for conformational changes while docking atomistic structures into cryo-EM maps is flexible fitting. Such a procedure is especially necessary if the conformational state captured by the EM map is significantly different than of the atomistic model used for fitting [1,105]. Most of the flexible fitting methods are deforming atomistic structures along the density of cryo-EM maps by using physics-based approaches, such as Normal Modes [106–108], MD [108–110] and simulated annealing protocols [111]. Normal Modes-based methods use a linear combination of low-frequency modes to morph the structure into the density map. MD-based techniques introduce a biasing potential, which forces the structure to fit into the EM map. SA methods translate the density information into a restraining term and minimize the resulting energy function. A recent review of Ahmed *et al.* [105] discussed whether there is a consensus among these fitting approaches. For this, one software package was selected for each of the different fitting method (Normal Modes: NMFF [112,113], MD: MDFit [114], SA: YUP.SCX [111]) and run on a benchmark of five large-sized proteins (350-800 amino acids). All cases had a medium resolution cryo-EM map (10-13 Å) that captured a different conformational state than that of the available structures. This comparison disclosed that, albeit the flexible fitting methods did differ, the resulting models were in consensus for the majority of cases and they could address collective conformational changes better than rigid fitting techniques (run with Situs [115]) [105].

In all of the above examples, only one type of data was used at a time. There are, however, a number of programs that can deal with various sources of information in a versatile manner to model large and dynamic macromolecular assemblies. A first example is RNABuilder, developed by Flores *et al.* [116]. RNABuilder offers an efficient way to deal with large-scale conformational changes: by using internal coordinates it reduces the number of degrees of freedom and treats the sub-domains of the RNA molecule as rigid bodies that are connected via flexible linkers. If present,

contact information, coming from NMR, cross-linking experiments, functional assays, bioinformatics, or any other source can be used to drive the folding [116]. RNABuilder was able to fold the 160 residue P4/P6 domain of a ribozyme to ~10Å away from the known crystal structure, 6Å lower RMSD than any of the previously published methods, and this in an order of magnitude shorter time [117]. Since its initial demonstration for RNA folding, it has been extended to other types of molecules in order to provide a cheap and generic solution to deal with large conformational changes (Samuel Flores, *personal communication*).

Another method, which is able to incorporate different types of data in a versatile way, is the Integrative Modeling Platform (IMP) developed in the Sali group. Based on the gathered experimental information and the chosen system representation, IMP first translates data into spatial restraints, which are then used to generate models by using various energy minimization functions and protocols. The key aspects of IMP are that it allows the inclusion of different data types, like contacts, proximity, distances, shape, etc., and resolution into its target function, and the simultaneous use of mixed (coarse- and fine-grained) system representation of a system [66]. The remarkable capability and efficiency of this kind of integrative approach was illustrated by the modeling of gigantic molecular machines, such as the Nuclear Pore Complex [118], the eukaryotic Ribosome [119] and, more recently, the 26S Proteasome [120]. In order to model the latter, information extracted from cryo-EM, X-Ray crystallography, chemical cross-linking experiments and bioinformatics approaches, like comparative modeling, was integrated. All these data were translated into spatial restraints that were applied in various combinations during consecutive modeling steps, consisting of subunit localization, fitting (with Multifit [104]), flexible refinement and clustering. This work shed light onto the macromolecular arrangement within the 26S proteasome and provided significant insights into the sequence of events prior to degradation [120].

In two other recent examples of integrative modeling, Campos *et al*. [121] and Loquet *et al*. [122] were able to model the pilus/needle of the bacterial secretion systems (pilus of the Type II and Type IV [121], needle of the Type III [122]). Instead of fitting the individual promoters into the available medium-to-low resolution cryo-EM maps in a rigid manner, they extracted biophysical properties from the EM maps to guide the search, rather than using the maps themselves, and allowed flexibility during their modeling.

Campos *et al*. used data from mutation experiments (salt-bridge charge inversion, double cysteine substitution and cross-linking), together with the biophysical information taken from low-resolution EM maps (symmetry of the assembly, degree of rise and angle per subunit), in order to model the Type II (from *Klebsiella oxytoca* and *Vibrio cholerae*) and Type IV (from *Neisseria gonorrhoeae*) pilus. [121] Through

starting with pili protomers and imposing a multi-stage minimization protocol followed by clustering and a final MD refinement step, they could model the pilus at atomistic detail and even suggest its handedness based on the number of restraint violations. Loquet *et al.*, on the other hand, started from extended polypeptide chains of the protomers and applied the *fold-and-dock* protocol [123] of Rosetta, in order to model the Type III (from *Salmonella typhimurium*) secretion needle. During modeling they enforced restraints translated from solid state NMR (chemical shifts, inter- and intra- subunit distances), STEM (axial properties) and cryo-EM (radius of the needle) experiments). They could not distinguish the handedness of the needle, but could reveal that having 11 subunits per two turns (rather than 9, 13 or 15) resulted in the least violations. These two recent examples of integrative modeling, which provided atomistic insight into different kinds of bacterial secretion systems, can easily be extended to model other supramolecular assemblies with helical symmetry.

A final but not least example is the Inferential Structure Determination (ISD) framework, which was developed as a NMR structure calculation suite [124,125]. ISD employs an unconventional but correct way to deal with sparse and imperfect data: it makes use of tools and concepts from Bayesian theory [125]. The relevance of an experimental observation for structure calculation is assessed rather than taking its contribution for granted. To do so, ISD follows: (i) Thorough exploration of the conformational space by replica-exchange Monte Carlo, (ii) calculation of occurrence frequencies of protein conformations that are compatible with the available data and (iii) translation of those frequencies into likelihoods toward inferring what is the "critical" set of information to calculate the structure and the degree of its significance (weight). With this statistical approach, ISD eliminates any bias introduced by empirical weights used in conventional structure calculation methods [125,126]. Recently this Bayesian approach has been extended to enhance the resolution of intermediate- and low-resolution cryo-EM density maps [127].

## 4. High Ambiguity Data-Driven Docking: HADDOCK

Our in house, information-driven macromolecular docking program, HADDOCK [65,128], is another example of an integrative approach for the modeling of biomolecular complexes. HADDOCK allows the inclusion of (sparse) data coming from various experimental sources and can deal simultaneously with molecules of different nature, i.e. proteins, peptides, small molecules, oligosaccharides, RNA, DNA [65,128]. The docking procedure is composed of three stages: (i) initial docking by rigid body energy minimization (*it0*), (ii) semi-flexible refinement in torsion angle space (*it1*) and (iii) final refinement in explicit solvent (*water*). The binding mode of the complex is roughly determined during *it0* and then a pre-defined percentage of

the top-ranking solutions according to the HADDOCK score (a weighted sum of electrostatics, van der Waals, restraint energies, buried surface area and an empirical desolvation term [129]), are selected for further refinement. The consecutive refinement steps allow for small- to medium-range conformational changes while improving the overall score of the models (**Figure 2**). The final structures are clustered based on their pairwise ligand interface RMSD and the average cluster scores are calculated over the top 4 members of each cluster [65,128].

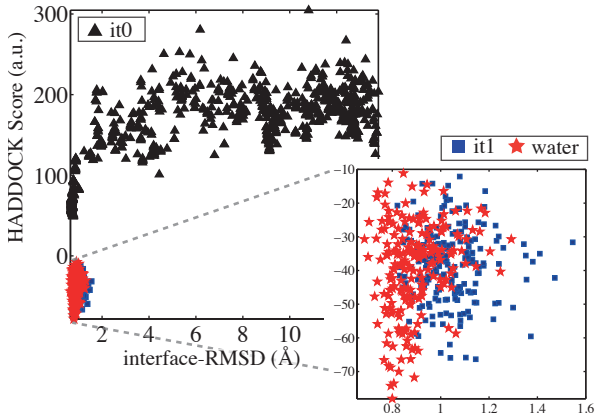



**Figure 2. HADDOCK score as a function of the interface-RMSD for models generated at the various stages of the protocol.** Solutions obtained after the initial rigid body energy minimization are indicated by black triangles. These are scored and the top 200-400 hundreds are selected for semi-flexible refinement in torsion angle space (blue squares) (*it1*) followed by a final explicit solvent refinement (*water*) (red stars). As represented in the inset, flexible refinement brings the models towards a lower energy state (resulting in lower interface-RMSDs and HADDOCK scores).

HADDOCK was originally developed to make use of NMR data, and in particular of chemical shift perturbation data (see Section 2.1.), but throughout the years it has extended its capacity markedly [65,128]. Currently it can translate most of the information sources listed in **Table 1** into structural restraints (or additional scoring term), except for cryo-EM data, although work in this direction is ongoing. HADDOCK uses a flat-bottom, "soft-square" potential to impose restraints [79]. This potential behaves harmonically up to violations of 2Å, after which it switches smoothly to a linear one. Such a modification avoids enormous forces due to large violations that can result in instabilities of the calculations (see **Section 2.2.**) [71,77]. The flat-bottom potential, enables the incorporation of restraints with upper and lower limits to account for the uncertainty of the measurements. Information about interfaces (but not the specific contacts made) is converted into Ambiguous

Interaction Restraints (AIRs). AIRs are composed of *active* (residues that are known to make contact) and *passive* (residues that potentially make contact – usually the surface neighbors of *active*'s) residues. Those residues are used to define a network of ambiguous distance restraints, which ensures that an *active* residue on the surface of a biomolecule should be in close vicinity to any *active* or *passive* residues on the partner biomolecule. If the list of interacting residues is not very accurate, e.g. in the case of bioinformatics predictions, then a user-defined percentage of the restraints can be discarded at random during docking and refinement (50% by default). Another key advantage of HADDOCK is its flexibility in imposing the restraints: users protocol and can change the weights assigned to each of them depending on the data accuracy and confidence in the data. All of these features are also offered via HADDOCK's user-friendly web server interface [128] at:

**http://haddock.science.uu.nl/**

## 5. Scope of the thesis

The comprehensive introduction presented above illustrated the current state of integrative approaches and the challenges of the field. Within this context, in this thesis, we focus on our information-driven docking program HADDOCK, pushing back the limits of its applicability in integrative modeling. To that end, in **Chapter 2**, a method to model generic multi-body complexes by simultaneous docking of all components is described and its performance is depicted for six multimer complexes, composed of five symmetric protein homo-oligomers and one symmetric protein–DNA complex. **Chapter 3** discusses an efficient divide-and-conquer approach, built on the multi-body docking ability of HADDOCK described in **Chapter 2**, to deal with large conformational changes upon binding. The performance of this approach is benchmarked on a set of 11 dimeric protein–protein complexes, covering a vast range of conformational change from 1.5 Å to as much as 19.5 Å! In **Chapter 4**, the usefulness of low-resolution shape data for scoring docking decoys is explored. A new scoring function is introduced that combines the regular HADDOCK score with the low-resolution shape information from either Small Angle X-ray Scattering or Collision Cross Section data. In **Chapter 5**, the methods developed in this thesis are demonstrated on two challenging real-case examples, for which different types of experimental data area available. In the first example, the structural basis for the ubiquitin linkage selectivity of a deubiquitination enzyme is investigated in the light of NMR and mutagenesis data. In the second one, a complex capturing the transfer of histones H3-H4 from one histone-chaperone to another one is modeled based on SAXS data. Finally, the thesis ends with a **Perspectives** section on the future and limitations of integrative modeling.

## References

[1] G.C. Lander, H.R. Saibil, E. Nogales, Go hybrid: EM, crystallography, and beyond, Curr. Opin. Struct. Biol. 22 (2012) 627–635.
[2] R.B. Russell, F. Alber, P. Aloy, F.P. Davis, D. Korkin, M. Pichaud, et al., A structural perspective on protein-protein interactions, Curr. Opin. Struct. Biol. 14 (2004) 313–324.
[3] C.L. Lawson, M.L. Baker, C. Best, C. Bi, M. Dougherty, P. Feng, et al., EMDataBank.org:
[4] C. Yang, G. Ji, H. Liu, K. Zhang, G. Liu, F. Sun, et al., Cryo-EM structure of a transcribing cypovirus, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 6118–6123.
[5] X. Zhang, S. Sun, Y. Xiang, J. Wong, T. Klose, D. Raoult, et al., Structure of Sputnik, a virophage, at 3.5-A resolution, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 18431–18436.
[6] T.F. Lerch, J.K. O'Donnell, N.L. Meyer, Q. Xie, K.A. Taylor, S.M. Stagg, et al., Structure of AAV-DJ, a retargeted gene therapy vector: cryo-electron microscopy at 4.5 Å resolution, Structure. 20 (2012) 1310–1320.
[7] A. Stein, R. Mosca, P. Aloy, Three-dimensional modeling of protein interactions and complexes is going 'omics, Curr. Opin. Struct. Biol. 21 (2011) 200–208.
[8] A. Melquiond, E. Karaca, P.L. Kastritis, A. Bonvin, Next challenges in protein-protein docking: from proteome to interactome and beyond, WIREs Comput Mol Sci. 2 (2011) 642–651.
[9] N.P. Cowieson, B. Kobe, J.L. Martin, United we stand: combining structural methods, Curr. Opin. Struct. Biol. 18 (2008) 617–622.
[10] D. Muradov, B. Kobe, E.N. Dixon, T. Huber, Hybrid Methods for Protein Structure Prediction, in: Hybrid Methods for Protein Structure Prediction, John Wiley & Sons, 2010: pp. 265–277.
[11] F. Alber, F. Förster, D. Korkin, M. Topf, A. Sali, Integrating diverse data for structure determination of macromolecular assemblies, Annu. Rev. Biochem. 77 (2008) 443–477.
[12] A.S.J. Melquiond, A.M.J.J. Bonvin, Data-driven docking: using external information to spark the biomolecular rendez-vous, in: Protein-Protein Complexes: Analysis, Modeling and Drug Design, Imperial College Press, 2010: pp. 183–209.
[13] M.F. Lensink, S.J. Wodak, Blind predictions of protein interfaces by docking calculations in CAPRI, Proteins: Struct. Funct. Bioinf. 78 (2010) 3085–3095.
[14] P.L. Kastritis, A.M.J.J. Bonvin, Are Scoring Functions in Protein-Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark, J Proteome Res. 9 (2010) 2216–2225.
[15] B. Simon, T. Madl, C.D. Mackereth, M. Nilges, M. Sattler, An efficient protocol for NMR-spectroscopy-based structure determination of protein complexes in solution, Angew. Chem., Int. Ed. 49 (2010) 1967–1970.
[16] D.K. Clare, D. Vasishtan, S. Stagg, J. Quispe, G.W. Farr, M. Topf, et al., ATP-triggered conformational changes delineate substrate-binding and -folding mechanics of the GroEL chaperonin, Cell. 149 (2012) 113–123.
[17] T. Madl, F. Gabel, M. Sattler, NMR and small-angle scattering-based structural analysis of protein complexes in solution, J. Struct. Biol. 173 (2011) 472–482.
[18] S. Kühner, V. van Noort, M.J. Betts, A. Leo-Macias, C. Batisse, M. Rode, et al., Proteome organization in a genome-reduced bacterium, Science. 326 (2009) 1235–1240.
[19] X. Wang, H.-W. Lee, Y. Liu, J.H. Prestegard, Structural NMR of protein oligomers using hybrid methods, J. Struct. Biol. 173 (2011) 515–529.
[20] M.A. McCoy, D.F. Wyss, Structures of Protein–Protein Complexes Are Docked Using Only NMR Restraints from Residual Dipolar Coupling and Chemical Shift Perturbations, J. Am. Chem. Soc. 124 (2002) 2104–2105.

[21] S. McKenna, T. Moraes, L. Pastushok, C. Ptak, W. Xiao, L. Spyracopoulos, et al., An NMR-based model of the ubiquitin-bound human ubiquitin conjugation complex Mms2.Ubc13. The structural basis for lysine 63 chain catalysis, J. Biol. Chem. 278 (2003) 13151–13158.
[22] K.J. Walters, P.J. Lech, A.M. Goh, Q. Wang, P.M. Howley, DNA-repair protein hHR23a alters its protein structure upon binding proteasomal subunit S5a, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 12694–12699.
[23] S.D. Emerson, R. Palermo, C.-M. Liu, J.W. Tilley, L. Chen, W. Danho, et al., NMR characterization of interleukin-2 in complexes with the IL-2Ralpha receptor component, and 12 (2003) 811–22.
[24] J.G. Mandell, A.M. Falick, E.A. Komives, Identification of protein-protein interfaces by decreased amide proton solvent accessibility, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 14705–14710.
[25] G. Pintacuda, G. Otting, Identification of protein surfaces by NMR measurements with a
[26] A. Mohd-Sarip, J.A. van der Knaap, C. Wyman, R. Kanaar, P. Schedl, C.P. Verrijzer, Architecture of a polycomb nucleoprotein complex, Mol. Cell. 24 (2006) 91–100.
[27] T. Clackson, J.A. Wells, A hot spot of binding energy in a hormone-receptor interface, Science. 267 (1995) 383–386.
[28] W.S. Valdar, J.M. Thornton, Conservation helps to identify biologically relevant crystal contacts, J. Mol. Biol. 313 (2001) 399–416.
[29] Q.C. Zhang, D. Petrey, R. Norel, B.H. Honig, Protein interface conservation across structure space, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 10896–901.
[30] F. Pazos, M. Helmer-Citterich, G. Ausiello, A. Valencia, Correlated mutations contain information about protein-protein interaction, J. Mol. Biol. 271 (1997) 511–523.
[31] I. Kufareva, L. Budagyan, E. Raush, M. Totrov, R. Abagyan, PIER: protein interface recognition for structural proteomics, Proteins: Struct. Funct. Bioinf. 67 (2007) 400–417.
[32] J. Garcia-Garcia, J. Bonet, E. Guney, O. Fornes, J. Planas, B. Oliva, Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details, Mol. Inf. 31 (2012) 342–362.
[33] E. Karaca, A.S.J. Melquiond, S.J. de Vries, P.L. Kastritis, A.M.J.J. Bonvin, Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server, Mol. Cell. Proteomics. 9 (2010) 1784–1794.
[34] G. Otting, K. Wüthrich, Heteronuclear filters in two-dimensional [1H, 1H]-NMR spectroscopy: combined use with isotope labelling for studies of macromolecular conformation and intermolecular interactions, Q. Rev. Biophys. 23 (1990) 39–96.
[35] G.M. Clore, Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear Overhauser enhancement data and dipolar couplings by rigid body minimization, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 9021–9025.
[36] E.G. Stein, L.M. Rice, A.T. Brünger, Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation, J. Magn. Reson. 124 (1997) 154–164.
[37] M. Trester-Zedlitz, K. Kamada, S.K. Burley, D. Fenyö, B.T. Chait, T.W. Muir, A modular cross-linking approach for exploring protein interactions, J. Am. Chem. Soc. 125 (2003) 2416–2425.
[38] M.M. Young, N. Tang, J.C. Hempel, C.M. Oshiro, E.W. Taylor, I.D. Kuntz, et al., High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 5802–586.
[39] A. Leitner, T. Walzthoeni, A. Kahraman, F. Herzog, O. Rinner, M. Beck, et al., Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics, Mol. Cell. Proteomics. 9 (2010) 1634–1649.
[40] J. Rappsilber, The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes, J. Struct. Biol. 173 (2011) 530–540.

[41] I. Bertini, C. Luchinat, G. Parigi, Magnetic susceptibility in paramagnetic NMR, Prog. Nucl. Magn. Reson. Spectrosc. 40 (2002) 249–273.
[42] G.F. White, L. Ottignon, T. Georgiou, C. Kleanthous, G.R. Moore, A.J. Thomson, et al., Analysis of nitroxide spin label motion in a protein-protein complex using multiple frequency EPR spectroscopy, J. Magn. Reson. 185 (2007) 191–203.
[43] H.J. Steinhoff, Inter- and intra-molecular distances determined by EPR spectroscopy and site-directed spin labeling reveal protein-protein and protein-oligonucleotide interaction, Biol.
[44] G. Otting, Protein NMR using paramagnetic ions, Annu. Rev. Biophys. 39 (2010) 387–405.
[45] A.T. Brunger, P. Strop, M. Vrljic, S. Chu, K.R. Weninger, Three-dimensional molecular modeling with single molecule FRET. J. Struct. Biol. 173 (2011) 497–505.
[46] A. Cha, G.E. Snyder, P.R. Selvin, F. Bezanilla, Atomic scale movement of the voltage-sensing region in a potassium channel measured via spectroscopy, Nature. 402 (1999) 809–813.
[47] T. Ha, Ligand-induced conformational changes observed in single RNA molecules, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 9077–9082.
[48] A. Bax, G. Kontaxis, N. Tjandra, Dipolar couplings in macromolecular structure determination, Methods Enzym. 339 (2001) 127–174.
[49] R. Brüschweiler, X. Liao, P.E. Wright, Long-range motional restrictions in a multidomain zinc-finger protein from anisotropic tumbling, Science. 268 (1995) 886–889.
[50] A. Grishaev, J. Wu, J. Trewhella, A. Bax, Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data, J. Am. Chem. Soc. 127 (2005) 16621–16628.
[51] F. Gabel, B. Simon, M. Nilges, M. Petoukhov, D. Svergun, M. Sattler, A structure refinement protocol combining NMR residual dipolar couplings and small angle scattering restraints, J. Biomol. NMR. 41 (2008) 199–208.
[52] C.D. Putnam, M. Hammel, G.L. Hura, J.A. Tainer, X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution, Q. Rev. Biophys. 40 (2007) 191–285.
[53] J. Kuszewski, A.M. Gronenborn, G.M. Clore, Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration, J. Am. Chem. Soc. 121 (1999) 2337–2338.
[54] A. Fotin, Y. Cheng, N. Grigorieff, T. Walz, S.C. Harrison, T. Kirchhausen, Structure of an auxilin-bound clathrin coat and its implications for the mechanism of uncoating, Nature. 432 (2004) 649–653.
[55] E. Jurneczko, P.E. Barran, How useful is ion mobility mass spectrometry for structural biology? The relationship between protein crystal structures and their collision cross sections in the gas phase, Analyst. 136 (2011) 20–28.
[56] C. Uetrecht, I.M. Barbu, G.K. Shoemaker, E. van Duijn, A.J. Heck, Interrogating viral capsid assembly with ion mobility–mass spectrometry, Nat. Chem. 3 (2010) 126–132.
[57] C. Uetrecht, R.J. Rose, E. van Duijn, K. Lorenzen, A.J. Heck, Ion mobility mass spectrometry of proteins and protein assemblies, Chem. Soc. Rev. 39 (2010) 1633–1655.
[58] I. Halperin, B. Ma, H. Wolfson, R. Nussinov, Principles of docking: An overview of search algorithms and a guide to scoring functions, Proteins: Struct. Funct. Bioinf. 47 (2002) 409–443.
[59] C. Pons, S. Grosdidier, A. Solernou, L. Pérez-Cano, J. Fernández-Recio, Present and future challenges and limitations in protein-protein docking, Proteins: Struct. Funct. Bioinf. 78 (2010) 95–108.
[60] I.S. Moreira, P.A. Fernandes, M.J. Ramos, Protein-protein docking dealing with the unknown, J. Comput. Chem. 31 (2010) 317–342.
[61] A. Solernou, J. Fernandez-Recio, pyDockCG: new coarse-grained potential for protein-protein docking, J. Phys. Chem. B. 115 (2011) 6032–6039.

[62] D.W. Ritchie, G.J. Kemp, Protein docking using spherical polar Fourier correlations, Proteins: Struct. Funct. Bioinf. 39 (2000) 178–194.
[63] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, I.A. Vakser, Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 2195–2199.
[64] E. Mashiach, D. Schneidman-Duhovny, A. Peri, Y. Shavit, R. Nussinov, H.J. Wolfson, An 3204.
[65] C. Dominguez, R. Boelens, A.M.J.J. Bonvin, HADDOCK: a protein-protein docking approach based on biochemical or biophysical information, J. Am. Chem. Soc. 125 (2003) 1731–1737.
[66] D. Russel, K. Lasker, B. Webb, J. Velázquez-Muriel, E. Tjioe, D. Schneidman-Duhovny, et al, Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies, PLoS Biol. 10 (2012) e1001244.
[67] J.J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, et al, Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations, J. Mol. Biol. 331 (2003) 281–299.
[68] W. Braun, N. Go, Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm, J. Mol. Biol. 186 (1985) 611–626.
[69] L. Verlet, Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules, Phys. Rev. 159 (1967) 98–103.
[70] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of State Calculations by Fast Computing Machines, J. Chem. Phys. 21 (1953) 1087.
[71] A.T. Brunger, P.D. Adams, L.M. Rice, Annealing in crystallography: a powerful optimization tool, Prog. Biophys. Mol. Biol. 72 (1999) 135–155.
[72] A.D.J. van Dijk, R. Boelens, A.M.J.J. Bonvin, Data-driven docking for the study of biomolecular complexes, FEBS J. 272 (2005) 293–312.
[73] Andrew R. Leach, Constraint Dynamics, in: Molecular Modelling: Principles and Applications, 2nd ed, Pearson Education, Dorchester, 2001: pp. 368–369.
[74] A.T. Brunger, Version 1.2 of the Crystallography and NMR system, Nat. Protoc. 2 (2007) 2728–2733.
[75] W.A. Hendrickson, Stereochemically restrained refinement of macromolecular structures, Methods Enzym. 115 (1985) 252–270.
[76] G.M. Clore, M. Nilges, D.K. Sukumaran, A.T. Brünger, M. Karplus, A.M. Gronenborn, The three-dimensional structure of alpha1-purothionin in solution: combined use of nuclear magnetic resonance, distance geometry and restrained molecular dynamics, EMBO J. 5 (1986) 2729–2735.
[77] M. Nilges, Structure calculation from NMR data, Curr. Opin. Struct. Biol. 6 (1996) 617–623.
[78] P.D. Adams, N.S. Pannu, R.J. Read, A.T. Brunger, Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 5018–5023.
[79] M. Nilges, A.M. Gronenborn, A.T. Brünger, G.M. Clore, Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2, Protein Eng. 2 (1988) 27–38.
[80] J.P. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca, A.S. Melquiond, et al., Clustering biomolecular complexes by residue contacts similarity, Proteins: Struct. Funct. Bioinf. 80 (2012) 1810–1817.
[81] E. Feliu, B. Oliva, How different from random are docking predictions when ranked by scoring functions?, Proteins: Struct. Funct. Bioinf. 78 (2010) 3376–3385.

[82] C.M. Pickart, R.E. Cohen, Proteasomes and their kin: proteases in the machine age, Nat. Rev. Mol. Cell Biol. 5 (2004) 177–187.
[83] K.Y. Sanbonmatsu, Computational studies of molecular machines: the ribosome, Curr. Opin. Struct. Biol. 22 (2012) 168–174.
very large macromolecular assemblies, Curr. Opin. Struct. Biol. 17 (2007) 572–579.
[85] W.F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, et al., Biomolecular modeling: Goals, problems, perspectives, Angew. Chem. Int. Ed. 45 (2006) 4064–4092.
[86] Y. Zhang, Progress and challenges in protein structure prediction, Curr. Opin. Struct. Biol. 18 (2008) 342–348.
[87] M. Christen, W.F. van Gunsteren, On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review, J. Comput. Chem. 29 (2008) 157–166.
[88] V. Tozzini, Coarse-grained models for proteins, Curr. Opin. Struct. Biol. 15 (2005) 144–150.
[89] M. Müller, K. Katsov, M. Schick, Coarse-grained models and collective phenomena in membranes: Computer simulation of membrane fusion, J. Polym. Sci. B Polym. Phys. 41 (2003) 1441–1450.
[90] I. André, P. Bradley, C. Wang, D. Baker, Prediction of the structure of symmetrical protein assemblies, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 17656–17661.
[91] E. Mashiach-Farkash, R. Nussinov, H.J. Wolfson, SymmRef: A flexible refinement method for symmetric multimers, Proteins: Struct. Funct. Bioinf. 79 (2011) 2607-2623.
[92] A.M.J.J. Bonvin, Flexible protein-protein docking, Curr. Opin. Struct. Biol. 16 (2006) 194–200.
[93] S.E. Dobbins, V.I. Lesk, M.J.E. Sternberg, Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 10390–10395.
[94] M. Zacharias, Accounting for conformational changes during protein-protein docking, Curr. Opin. Struct. Biol. 20 (2010) 180–186.
[95] E. Karaca, A.M.J.J. Bonvin, A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes, Structure. 19 (2011) 555-565.
[96] S. Chaudhury, J.J. Gray, Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles, J. Mol. Biol. 381 (2008) 1068–1087.
[97] M. Król, R.A.G. Chaleil, A.L. Tournier, P.A. Bates, Implicit flexibility in protein docking: cross-docking and local refinement, Proteins: Struct. Funct. Bioinf. 69 (2007) 750–757.
[98] A.C. Steven, W. Baumeister, The future is hybrid, J. Struct. Biol. 163 (2008) 186–195.
[99] H.M. Berman, G.J. Kleywegt, H. Nakamura, J.L. Markley, The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future, Structure. 20 (2012) 391–396.
[100] M.F. Lensink, R. Méndez, S.J. Wodak, Docking and scoring protein complexes: CAPRI 3rd Edition, Proteins: Struct. Funct. Bioinf. 69 (2007) 704–718.
[101] C. Pons, M. D'Abramo, D.I. Svergun, M. Orozco, P. Bernadó, J. Fernández-Recio, Structural characterization of protein-protein complexes by integrating computational docking with small-angle scattering data, J. Mol. Biol. 403 (2010) 217–230.
[102] D. Schneidman-Duhovny, M. Hammel, A. Sali, Macromolecular Docking Restrained by a Small Angle X-Ray Scattering Profile, J. Struct. Biol. 173 (2010) 461-471.
[103] C. Schmitz, A.S.J. Melquiond, S.J. de Vries, E. Karaca, M. van Dijk, P.L. Kastritis, et al., Protein-Protein Docking with HADDOCK, in: I. Bertini, K.S. McGreevy, G. Parigi (Eds.), NMR of Biomolecules: Towards Mechanistic Systems Biology, 1st ed, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2012: pp. 521–535.

[104] K. Lasker, A. Sali, H.J. Wolfson, Determining macromolecular assembly structures by molecular docking and fitting into an electron density map, Proteins: Struct. Funct. Bioinf. 78 (2010) 3205–3211.
[105] A. Ahmed, P.C. Whitford, K.Y. Sanbonmatsu, F. Tama, Consensus among flexible fitting approaches improves the interpretation of cryo-EM data,J. Struct. Biol. 177 (2012) 561–70.
[106] M. Delarue, P. Dumas, On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 6957–6962.
of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase, Biophysical Journal. 88 (2005) 818–827.
[108] K. Suhre, J. Navaza, Y.H. Sanejouand, NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps, Acta Crystallogr. Sect. D: Biol. Crystallogr. 62 (2006) 1098–1100.
[109] K.-Y. Chan, L.G. Trabuco, E. Schreiner, K. Schulten, Cryo-electron microscopy modeling by the molecular dynamics flexible fitting method, Biopolymers. 97 (2012) 678–686.
[110] A.H. Ratje, J. Loerke, A. Mikolajka, M. Brünner, P.W. Hildebrand, A.L. Starosta, et al., Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites, Nature. 468 (2010) 713–716.
[111] R.K.-Z. Tan, B. Devkota, S.C. Harvey, YUP.SCX: coaxing atomic models into medium resolution electron density maps, J. Struct. Biol. 163 (2008) 163–174.
[112] F. Tama, O. Miyashita, C.L. Brooks, Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM, J. Struct. Biol. 147 (2004) 315–26.
[113] F. Tama, O. Miyashita, C.L. Brooks, Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis, J. Mol. Biol. 337 (2004) 985–99.
[114] P.C. Whitford, A. Ahmed, Y. Yu, S.P. Hennelly, F. Tama, C.M.T. Spahn, et al., Excited states of ribosome translocation revealed through integrative molecular modeling, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 18943–18948.
[115] W. Wriggers, R.A. Milligan, J.A. McCammon, Situs: A package for docking crystal structures into low-resolution maps from electron microscopy, J. Struct. Biol. 125 (1999) 185–195.
[116] S.C. Flores, M.A. Sherman, C.M. Bruns, P. Eastman, R.B. Altman, Fast flexible modeling of RNA structure using internal coordinates, IEEE/ACM Trans. Comput. Biol. Bioinf. / IEEE, ACM. 8 (2011) 1247–1257.
[117] S.C. Flores, R.B. Altman, Turning limited experimental information into 3D models of RNA, RNA. 16 (2010) 1769–1778.
[118] F. Alber, S. Dokudovskaya, L.M. Veenhoff, W. Zhang, J. Kipper, D. Devos, et al., Determining the architectures of macromolecular assemblies, Nature. 450 (2007) 683–94.
[119] D.J. Taylor, B. Devkota, A.D. Huang, M. Topf, E. Narayanan, A. Sali, et al., Comprehensive molecular structure of the eukaryotic ribosome, Structure. 17 (2009) 1591–1604.
[120] K. Lasker, F. Förster, S. Bohn, T. Walzthoeni, E. Villa, P. Unverdorben, et al., Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 1380–1387.
[121] M. Campos, O. Francetic, M. Nilges, Modeling pilus structures from sparse data, J. Struct. Biol. 173 (2011) 436–444.
[122] A. Loquet, N.G. Sgourakis, R. Gupta, K. Giller, D. Riedel, C. Goosmann, et al, Atomic model of the type III secretion system needle, Nature. 486 (2012) 276–279.

[123] R. Das, I. André, Y. Shen, Y. Wu, A. Lemak, S. Bansal, et al., Simultaneous prediction of protein folding and docking at high resolution, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 18978–18983.
[124] W. Rieping, M. Nilges, M. Habeck, ISD: a software package for Bayesian NMR structure calculation, Bioinformatics. 24 (2008) 1104–1105.
[125] M. Habeck, Statistical mechanics analysis of sparse data, J. Struct. Biol. 173 (2011) 541–548.
macromolecular structure determination, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 1756–
[127] M. Hirsch, B. Schölkopf, M. Habeck, A blind deconvolution approach for improving the resolution of cryo-EM density maps, J. Comput. Biol. 18 (2011) 335–346.
[128] S.J. de Vries, M. van Dijk, A.M.J.J. Bonvin, The HADDOCK web server for data-driven biomolecular docking, Nat. Protoc. 5 (2010) 883–897.
[129] J. Fernández-Recio, M. Totrov, R. Abagyan, Identification of protein-protein interaction sites from docking energy landscapes, J. Mol. Biol. 335 (2004) 843–865.
[130] H. Hwang, T. Vreven, J. Janin, Z. Weng, Protein-protein docking benchmark version 4.0, Proteins: Struct. Funct. Bioinf. 78 (2010) 3111–3114.
[131] G. Nicastro, S.V. Todi, E. Karaca, A.M.J.J. Bonvin, H.L. Paulson, A. Pastore, Understanding the Role of the Josephin Domain in the PolyUb Binding and Cleavage Properties of Ataxin-3, PLoS One. 5 (2010) e12430.
[132] J. Seebacher, P. Mallick, N. Zhang, J.S. Eddes, R. Aebersold, M.H. Gelb, Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing, J. Proteome Res. 5 (2006) 2270–2282.
[133] P. Aloy, R.B. Russell, Structural systems biology: modelling protein interactions, Nature Reviews. Mol. Cell Biology. 7 (2006) 188–197.
[134] H. Schreuder, C. Tardif, S. Trump-Kallmeyer, A. Soffientini, E. Sarubbi, A. Akeson, et al., A new cytokine-receptor binding mode revealed by the crystal structure of the IL-1 receptor with an antagonist, Nature. 386 (1997) 194–200.
[135] T. Kortemme, L.A. Joachimiak, A.N. Bullock, A.D. Schuler, B.L. Stoddard, D. Baker, Computational redesign of protein-protein interaction specificity, Nat. Struct. Mol. Biol. 11 (2004) 371–379.

Chapter

2

# Building Macromolecular Complexes by Information-driven Docking:

## Introducing the HADDOCK Multibody Server

Based on research article:
**Ezgi Karaca[*], Adrien S.J. Melquiond[*], Sjoerd J. de Vries[*], Panagiotis L. Kastritis, Alexandre M.J.J. Bonvin,** Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server**.** *Mol Cell Proteomics* 2010, 9: 1784–1794.
[*]These authors contributed equally to this work.

**Abstract**

Over the last years, large-scale proteomics studies have generated a wealth of information of biomolecular complexes. Adding the structural dimension to the resulting interactomes represents a major challenge that classical structural experimental methods alone will have difficulties to confront. To meet this challenge, complementary modeling techniques such as docking are thus needed. Among the current docking methods, HADDOCK (High Ambiguity Driven DOCKing) distinguishes itself from others by the use of experimental and/or bioinformatics data to drive the modeling process and has shown a strong performance in the critical assessment of prediction of interactions (CAPRI), a blind experiment for the prediction of interactions. Although most docking programs are limited to binary complexes, HADDOCK can deal with multiple molecules (up to six), a capability that will be required to build large macromolecular assemblies. We present here a novel web interface of HADDOCK that allows the user to dock up to six biomolecules simultaneously. This interface allows the inclusion of a large variety of both experimental and/or bioinformatics data and supports several types of cyclic

and dihedral symmetries in the docking of multibody assemblies. The server was tested on a benchmark of six cases, containing five symmetric homo-oligomeric protein complexes and one symmetric protein-DNA complex. Our results reveal that, in the presence of either bioinformatics and/or experimental data, HADDOCK shows an excellent performance: in all cases, HADDOCK was able to generate good to high quality solutions and ranked them at the top, demonstrating its ability to model symmetric multicomponent assemblies. Docking methods can thus play an important role in adding the structural dimension to interactomes. However, although the current docking methodologies were successful for a vast range of cases, considering the variety and complexity of macromolecular assemblies, inclusion of some kind of experimental information (e.g. from mass spectrometry, nuclear magnetic resonance, cryoelectron microscopy, etc.) will remain highly desirable to obtain reliable results.



## 1. Introduction

Proteins are the wheels and millstones of the complex machinery that underlies human life. Catalyzing a huge diversity of chemical processes, proteins work in close association with other biomolecules: nucleic acids, sugars, lipids, and other proteins. This huge network of protein interactions enables the cell to respond quickly to changes in the environment, such as temperature, oxygen, or nutrient concentration. However, to fully understand this network, insights at the atomic level are needed.

In the wake of the elucidation of the human genome [1, 2], many structural genomics projects are solving the structures of what is now becoming a considerable fraction of the human proteome [3]. These projects are now moving to the next level, which is solving the atomic resolution structures of protein complexes. However, this is a challenge that is considerably greater than obtaining the structures of single proteins. First of all, a protein can take part in 10 interactions on average; thus, the number of complexes is expected to be at least an order of magnitude larger than the proteome, and their composition can even vary over time. Second, associations between subunits in protein complexes are often weak and reversible, which make purification and crystallization difficult. Finally, there are some very well studied classes of interactions, such as enzyme-inhibitor, antibody-antigen, and GTPase-GAP (GTPase-activating protein) interactions, but these classes represent binary interactions between proteins. In contrast, many of the most important functions in the cell are carried out by large, dynamic molecular assemblies, such as the ribosome, the proteasome, the spliceosome, RNA polymerases, and the nuclear pore complex [4, 5]. For such assemblies, high-resolution methods such as X-Ray crystallography and NMR spectroscopy often provide atomic level information at the

level of individual subunits or subcomplexes, but they typically encounter difficulties at the level of the full complex.

Fortunately, low-resolution information about protein complexes can often be obtained. Affinity purification [6, 7] followed by mass spectrometry is a high throughput technique to study the composition of a complex. However, dissociation inside the mass spectrometer can be a problem for transient or unstable complexes in which case chemical cross-linking can help. Once the composition of the complex is known, there is a variety of experimental techniques available to obtain structural information on the complex. The most detailed information can be gathered by using data obtained from various NMR experiments, for example chemical shift perturbations [8] or residual dipolar couplings [9]; unfortunately, NMR is limited to complexes that are fairly small in size, making its applicability in the context of large assemblies less suited. Techniques that provide information about the shape of a protein complex, such as small angle X-Ray scattering (SAXS), cryoelectron tomography, and single molecule cryoelectron microscopy (cryo-EM), are more suited to characterize large complexes. Unfortunately, all of these techniques suffer from limitations in resolution that are either fundamental or caused by structural heterogeneities of the complex.

A well known approach to obtain information on residues at an interface is site-directed mutagenesis [10]. In principle, a loss of binding affinity indicates that the mutated residue mediates the interaction, although the reverse is not true. Also, one must take care of secondary effects, such as unfolding or conformational change caused by the mutation. Apart from that, very detailed information about interface residues can be obtained by extensive mutagenesis experiments, such as alanine scanning and double mutant cycles. Mass spectrometry offers the opportunity to get peptide level or residue level information about protein interfaces by accurate mass measurements of peptides from the protein complex, generated either a priori through proteolytic cleavage, or inside the mass spectrometer (MS/MS). For example, interface residues can be identified as residues that undergo slower hydrogen/deuterium exchange upon complex formation. This process can be monitored at the peptide level by mass spectrometry (or in smaller complexes, at the residue level by NMR), although this method is very sensitive to noise caused by conformational changes upon binding. In the same way, radical probe MS (RP-MS) uses differences in oxidation of residues by hydroxyl radicals generated in the mass spectrometer to identify interface residues. Finally, chemical cross-linking followed by MS can provide direct information about residue contact sites between different binding partners of the complex. Several cross-linking reagents can provide complementary information. However, it has been reported that the cross-linkers

may disrupt the structure of the protein complex and that care should therefore be taken to interpret the results [11].

There is a need for computational approaches to translate this low-resolution information into atomic resolution models that can provide functional and mechanistic insights. One of the most promising approaches is docking, the prediction of the structure of a complex starting from the free, unbound structures of its constituents. In recent years, docking methods have made much progress in the blind prediction of the structure of protein complexes as seen in the recent rounds of the critical assessment of prediction of interactions (CAPRI) experiment [12, 13]. Most docking methods are *ab initio*, which means that experimental data are not required. However, it is possible in several *ab initio* methods to use experimentally determined interface residues in the docking: in MolFit [14, 15] and ATTRACT [16, 17], it is possible to upweight the interaction scores of interface residues; in ZDOCK [18, 19], it is possible to block non-interface residues; and in PatchDock [20, 21], ZDOCK, pyDock [22, 23], and several other methods, it is possible to filter the docking results based on experimental information. Next to purely *ab initio* approaches, there are also methods that make use of different types experimental information, for example PROXIMO [24], based on RP-MS data, and MultiFit [25], a hybrid fitting/docking approach based on electron microscopy data.

A method that distinguishes itself from the variety of above mentioned docking approaches is HADDOCK [26–28]. In HADDOCK, the docking can be driven by a variety of experimental data using information about interface, contacts, and relative orientations inside a complex simultaneously. Originally developed for NMR data, HADDOCK is able to deal with a large variety of experimental data as shown in **Table 1**. Interface residues are defined as "*active residues*" that are believed to participate in the formation of the interface, and "*passive residues*" are those that are possibly at the interface; other kinds of data can be entered directly. (See the original HADDOCK studies [26–28] and **Materials and Methods** for more details.) HADDOCK has performed very well in translating these data into structures and structural models. More than 60 Protein Data Bank structures calculated using HADDOCK have been deposited to date as experimental structures in the Protein Data Bank [29]. Moreover, HADDOCK has shown a strong performance in CAPRI. Finally, HADDOCK is a general purpose program that can integrate many kinds of data, but even with a single source of data it is able to perform as well as more specialized programs: for example, HADDOCK was able to closely reproduce the NMR-calculated E2A-HPr complex using only chemical shift perturbation data. For the ribonuclease S-protein-peptide complex (Protein Data Bank code 1J80 [30]) for which RP-MS data are available, PROXIMO was able to closely reproduce the crystal structure (root mean square deviation (RMSD) of the top scoring model from the

reference crystal structure is 1.26 Å); using the same data, HADDOCK could get even closer with an RMSD of only 0.68 Å from the crystal structure (results not shown).

**Table 1. The various experimental data that can be incorporated in to HADDOCK**

| Experimental data | HADDOCK representation |
|---|---|
| Mutagenesis data | *Active* and *passive* residues |
| H/D exchange data | *Active* and *passive* residues |
| Bioinformatic interface predictions | *Active* and *passive* residues |
| Mass spectrometry data | |
| Cross-linking data | Custom CNS restraints |
| Radical probe mass spectrometry | *Active* and *passive* residues |
| Limited proteolysis mass spectrometry | *Active* and *passive* residues or directly, as an MTMDAT-generated HADDOCK parameter file |
| NMR data | |
| Chemical shift perturbation data | *Active* and *passive* residues |
| Cross-saturation experiments | *Active* and *passive* residues |
| Residual Dipolar Couplings | Directly |
| Diffusion anisotropy restraints | Directly |
| NOEs: as custom CNS restraints | Custom CNS restraints |
| Dihedral angles | Directly |
| Hydrogen bonds | Directly |
| Para-magnetic restraints | Under development |
| Shape data | |
| SAXS | Under development |
| EM | Under development |

Most docking methods are designed to deal with just two molecules, making their application limited with regard to large macromolecular assemblies. In most programs, multicomponent complexes can be assembled by adding each component one at a time, whereas simultaneous docking of the whole complex is typically not possible. Recently five *ab initio* docking programs (MolFit [31, 32], ClusPro [33], Rosetta [34], M-ZDOCK [35], and SymmDock [36]) gave birth to specific versions for the prediction of the symmetric multimers. Among these programs, MolFit, ClusPro, and Rosetta perform a rotational/translational search about the proper symmetry axes. These programs can deal with different types of cyclic and dihedral symmetries. Different than the other two, Rosetta is able to assemble complexes having helical and icosahedral symmetries. M-ZDOCK and SymmDock are suited for the prediction of macromolecules with cyclic symmetries. However, the ability to deal with arbitrary large molecular assemblies is currently rare. CombDock [37], which was

developed by the team of SymmDock, can build hetero-oligomer complexes, but it does not have a symmetry option. Only HADDOCK can deal with molecular complexes that are hetero-oligomers or homo-oligomers with arbitrary symmetry operators between and within each component.

The flexibility of HADDOCK comes at a price: it requires the user to have the structure calculation program CNS [38] installed and a considerable degree of expertise in its usage and molecular modeling in general, and it requires a cluster of computers. To alleviate this problem and to open up HADDOCK for a wide community, we have recently developed the HADDOCK web server [27]. The server offers multiple web interfaces, ranging from very simple and user-friendly to very powerful and flexible, exposing the full range of HADDOCK options to the expert user. However, up until now, the HADDOCK server was unable to deal with more than two molecules. Here we present a novel web interface for multibody docking of complexes. Like the HADDOCK program itself, the server supports the docking of up to six molecules simultaneously; all HADDOCK options, including symmetry restraints, are made available to the user. Even larger assemblies can in principle be modeled if the docking is performed in an incremental way. Here we demonstrate the performance of the multibody server on a small benchmark comprising complexes of various symmetries and increasing numbers of components (from three to five). To drive the docking, bioinformatics interface predictions and/or available experimental information were used. The HADDOCK server is available online at http://haddock.chem. uu.nl.

## 2. Materials and Methods

### *2.1. Ambiguous Interaction Restraints and Docking Protocol*

HADDOCK uses experimental and/or bioinformatics data to drive the complex formation *in silico*. The experimental and/or prediction data are used to define *active* and *passive* residues. *Active* residues are described as the identified interface residues, and *passive* residues correspond to their solvent-accessible neighbours. These are used to define a network of Ambiguous Interaction Restraints (AIRs) between the molecules to be docked. An AIR defines that a residue on the surface of a biomolecule should be in close vicinity to another residue or group of residues on the partner biomolecule when they form the complex. By default, this is described as an ambiguous distance restraint between all atoms of the source residue to all atoms of all target residue(s) that are assumed to be in the interface of the complex (**Figure 1**).

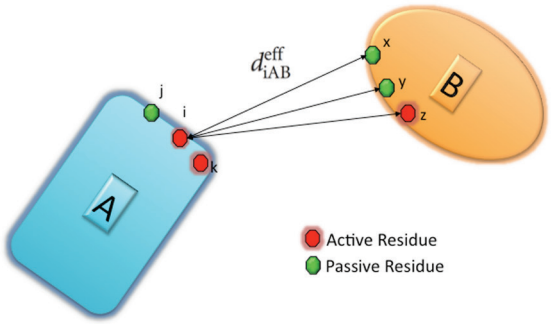



**Figure 1. Illustration of the ambiguous interaction restraints (AIRs) used in HADDOCK to drive the docking.** *Active residues* correspond to residue experimentally identified or predicted to be at the interface. *Passive residues* are surface neighbors of *active residue*s. AIRs are defined for each *active residue* with the effective distance being calculated from the sum of all individual distance between any atom of an *active residue* and any atom of all *active* and *passive residues* on the partner molecule (Eq. 1).

The effective distance between all those atoms, $d_{iAB}^{eff}$ is calculated as follows:

$$d_{iAB}^{eff} = \left( \sum_{m_{iA}=1}^{N_{Aatom}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{Batom}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-1/6} \qquad \text{(Eq. 1)}$$

Here $N_{Aatom}$ indicates all atoms of the source residue on molecule A, $N_{resB}$ the residues defined to be at the interface of the target molecule B and $N_{Batom}$ all atoms of a residue on molecule B. The $1/r^6$ summation somewhat mimics the attractive part of a Lennard-Jones potential and ensures that the AIRs are satisfied as soon as any two atoms of the biomolecules are in contact. The AIRs are incorporated as an additional energy term to the energy function that is minimized during the docking. The ambiguous nature of these restraints easily allows experimental data that often provide evidence for a residue making contacts to be used as a driving force for the docking. As such the AIRs define a network of restraints between the possible interaction interface(s) of the molecules to be docked without defining the relative orientation of the molecules, minimizing the necessary search through conformational space needed to assemble the interfaces.

The docking protocol in HADDOCK consists of three stages: (i) rigid body energy minimization (*it0*), (ii) semi-flexible refinement in torsion angle space (*it1*), and (iii) a final explicit solvent refinement (*water*). In the last two stages, flexible segments are typically defined automatically based on the identified intermolecular

contacts. The solutions are ranked at the end of each docking stage based on the following HADDOCK scoring functions:

**it0:** $0.01 \cdot E_{vdW} + 1.0 \cdot E_{Elec} + 0.01 \cdot E_{AIR} - 0.01 \cdot BSA + 1.0 \cdot E_{Desolv} + 0.1 \cdot E_{Sym}$ (Eq. 2)

**it1:** $1.0 \cdot E_{vdW} + 1.0 \cdot E_{Elec} + 0.1 \cdot E_{AIR} - 0.01 \cdot BSA + 1.0 \cdot E_{Desolv} + 0.1 \cdot E_{Sym}$ (Eq. 3)

**water:** $1.0 \cdot E_{vdW} + 0.2 \cdot E_{Elec} + 0.1 \cdot E_{AIR} + 1.0 \cdot E_{Desolv} + 0.1 \cdot E_{Sym}$ (Eq. 4)



The weighted parameters that are used in different stages of the scoring are: van der Waals ($E_{vdW}$), electrostatics ($E_{Elec}$), restraint violation ($E_{AIR}$), desolvation ($E_{Desolv}$) [39] symmetry restraint energies, and buried surface area (BSA). The solutions are clustered using a 7.5Å cut-off based on their pairwise RMSD values and the cluster ranks are determined according to the average energy of the four best structures of each cluster.

### *2.2. Dealing with symmetry*

HADDOCK can deal with biomolecules having cyclic (C2, C3, C5) symmetries or any combination thereof. This also allows dealing with dihedral symmetries since dihedral symmetry can be interpreted as a combination of cyclic symmetry pairs (e.g. D2 symmetry is a combination of 6 C2 symmetry pairs (see **Table 2**)). The symmetry restraints can be applied both within and between molecules. Compared to other docking programs supporting symmetric molecules, the unique characteristic of HADDOCK is that it applies symmetry on the molecules while docking them simultaneously. For the generation of symmetric complexes, two types of restraints should be used in combination: NCS (non-crystallographic symmetry) [40] and distance symmetry restraints [41, 42], both available within CNS.

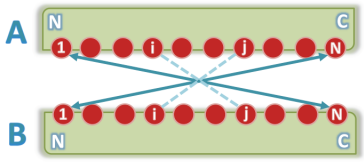
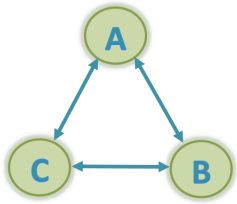
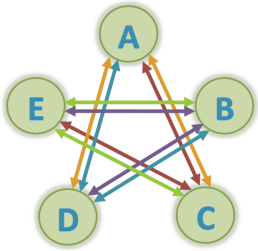
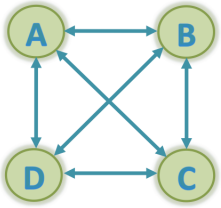
***Non-crystallographic symmetry.*** NCS restraints force two (or more) monomers to be identical without defining any symmetry operation between them. This is achieved through minimization of the following potential energy function:

$$E_{NCS} = k_{NCS} \sum_{a=1}^{A} \sum_{m=1}^{M} (x'_{am} - \bar{x}_a)^2 + \left(y'_{am} - \bar{y}_a\right)^2 + \left(z'_{am} - \bar{z}_a\right)^2 \qquad \text{(Eq. 5)}$$

This energy term is calculated after superposition of the monomers onto the first monomer. In the potential expression, A is the number of atoms, M is the number of monomers, $k_{NCS}$ is a constant, $(x'_{am}, y'_{am}, z'_{am})$ are the Cartesian coordinates of the $a^{th}$ atom on the $m^{th}$ monomer and $(\bar{x}_a, \bar{y}_a, \bar{z}_a)$ corresponds to the average position of $a^{th}$ atom with respect to the superimposed coordinates [41]. Using NCS restraints in HADDOCK only requires the user to define pairs of segments on which the NCS restraints will be applied. These can belong either to the same molecule or to

separate molecules allowing to define both intra- and intermolecular symmetries. The only requirement is that the number and type of atoms should be identical in both segments.

**Table 2. Definition and illustration of the symmetry restraining options in HADDOCK.**

| | | |
|---|---|---|
| C2 |  | $d(A_iB_j) = d(B_iA_j)$ |
| C3 |  | $d(AB) = d(BC)$<br>$d(BC) = d(CA)$<br>$d(CA) = d(AB)$ |
| C5 |  | $d(AC) = d(AD)$<br>$d(BD) = d(BE)$<br>$d(CE) = d(CA)$<br>$d(DA) = d(DB)$<br>$d(EB) = d(EC)$ |
| D2 |  | $d(AB) = d(BA)$<br>$d(AC) = d(CA)$<br>$d(AD) = d(DA)$<br>$d(BC) = d(CB)$<br>$d(BD) = d(DB)$<br>$d(CD) = d(DC)$ |

***Cyclic and Dihedral Symmetry.*** The implementation of this type of symmetry in HADDOCK is based on the symmetry distance restraints defined by Nilges [41, 42] for the NMR structure calculation of symmetrical dimers. The symmetry is imposed by requiring that pairs of intermolecular distances between all symmetric Cα atoms should have identical values. In the case of a dimer composed of molecules A and B, this condition can be illustrated as follows:

$$\Delta = d(A_i,B_j) - d(B_i,A_j) \quad \text{(Eq. 6)}$$

Δ is summed over all distances between $C_a$ atoms of the defined segments. Here the idea is to minimize Δ, so that the symmetric distances between the monomers are equal to each other. This is illustrated in **Table 2**. The major advantage of this approach is that it does not require knowledge of the position of the symmetry axis and it can be applied to different symmetries (C2, C3, C5 as shown in **Table 2**) and oligomeric proteins. The symmetric pairs should be defined as explained above for NCS restraints.

### *2.3. Docking Procedure for Symmetric Complexes*

All test cases, except for the protein-DNA complex, were docked using the multibody web interface of HADDOCK (**Table 3**). The procedure followed to dock the protein-DNA complex differs from the generic multibody docking protocol in the sense that two subsequent docking rounds were performed: in the second round custom-built DNA models that captured the conformational changes in the DNA from the first docking are used as starting structures. This approach allows modeling rather large deformations in the DNA and is explained comprehensively in a recent work of van *Dijk et al.* (M. van Dijk and A.M.J.J. Bonvin, 2010, *Nucleic Acids Res.*).

**Table 3. Properties of the Multimer Docking Benchmark**

| PDB ID | CATH Classification | Complex Type | Docking Type | Symmetry Type | # of amino acids |
|---|---|---|---|---|---|
| 1QU9 [66] | Mainly Beta | Homotrimer | Bound | C3 | 128 |
| 1URZ [67] | Mainly Alpha / | Homotrimer | Unbound | C3 | 400 |
| 1OUS [68] | Alpha Beta | Homotetramer | Bound | D2 | 114 |
| 1VIM [69] | Alpha Beta | Homotetramer | Bound | D2 | 200 |
| 1VPN [70] | Mainly Beta | Homopentamer | Bound | C5 | 289 |
| 3CRO [71] | Mainly Alpha | Homodimer-Double stranded | Unbound | C2 | 71 (Protein) |

In four of the test cases (Protein Data Bank codes 1QU9, 1OUS, 1VIM, and 1VPN), the interface information was obtained through the consensus interface prediction server CPORT using the “very sensitive” option. In the case of Protein Data Bank code 1URZ, a former CAPRI target, we used the same interface definition as was used previously in CAPRI [43]. For the protein-DNA complex, Protein Data Bank code 3CRO, sequence conservation and experimental data (mutagenesis and ethylation interference) were used to define the protein-DNA interaction site. The interface information was converted into AIRs via the setup page of the HADDOCK web site. The generated AIR files together with the input structures were then

supplied to the multibody server as an input for the docking. To favor compactness of the solution, center-of-mass restraints were enabled. For each complex, the proper combination of NCS and symmetry restraints were defined. Sampling of 180° rotated solutions was disabled. The number of structures was increased to 5000, 400, and 400 for it0, it1, and water, respectively. All other parameters were left at their default settings.



### *2.4. Evaluation of Docking Models*

The models were evaluated according to the CAPRI criteria [13]. For a complex to be classified as acceptable (one star), its interface root mean square deviation (i-RMSD) from the complex had to be lower than 4 Å, or its liga nd RMSD (l-RMSD) had to be lower than 10 Å. In addition, the fraction of native contacts (Fnat) had to be ≤0.1. For good predictions (two stars), the criteria were i-RMSD ≤2 Å or l-RMSD ≤5 Å and Fnat ≤0.3. For high quality predictions (three stars), the criteria were i-RMSD ≤1 Å or l-RMSD ≤1 Å and Fnat ≤0.5. A cluster was considered one/two/three star(s) if at least one of its top four members was of one-/two-/three-star quality or better.

## 3. Results

We have compiled a benchmark of six multimer assemblies. The complexes are homomeric with different numbers of components and various symmetries (see **Table 3**). One of them corresponds to a dimeric protein-DNA complex. In four cases, the docking was performed starting from the separated components of the crystal structure ("bound docking"). In one case (1URZ), the starting structures correspond to the dimeric form of the complex, whereas the trimeric form had to be predicted; this complex corresponds to a viral envelope protein that was a target in CAPRI (target 10). For the protein-DNA complex (3CRO), the docking was performed from the unbound conformation of the monomers and a canonical B-DNA model. In summary, our benchmark consists of four bound cases and two unbound cases.

For modeling of the benchmark complexes, we made use of the new multibody interface of the HADDOCK web server. The web server provides a user-friendly interface that gives full control over the various HADDOCK parameters and supports a wide range of experimental restraints (**Figure 2**). This interface is freely docking of up to six molecules and supports several types of cyclic symmetries (C2, C3, or C5) and any type of dihedral symmetry that can be expressed as a combination of the available cyclic symmetry pairs (see **Materials and Methods**). Our server is the first to support cyclic and dihedral symmetries at the same time and to allow simultaneous docking of up to six molecules.

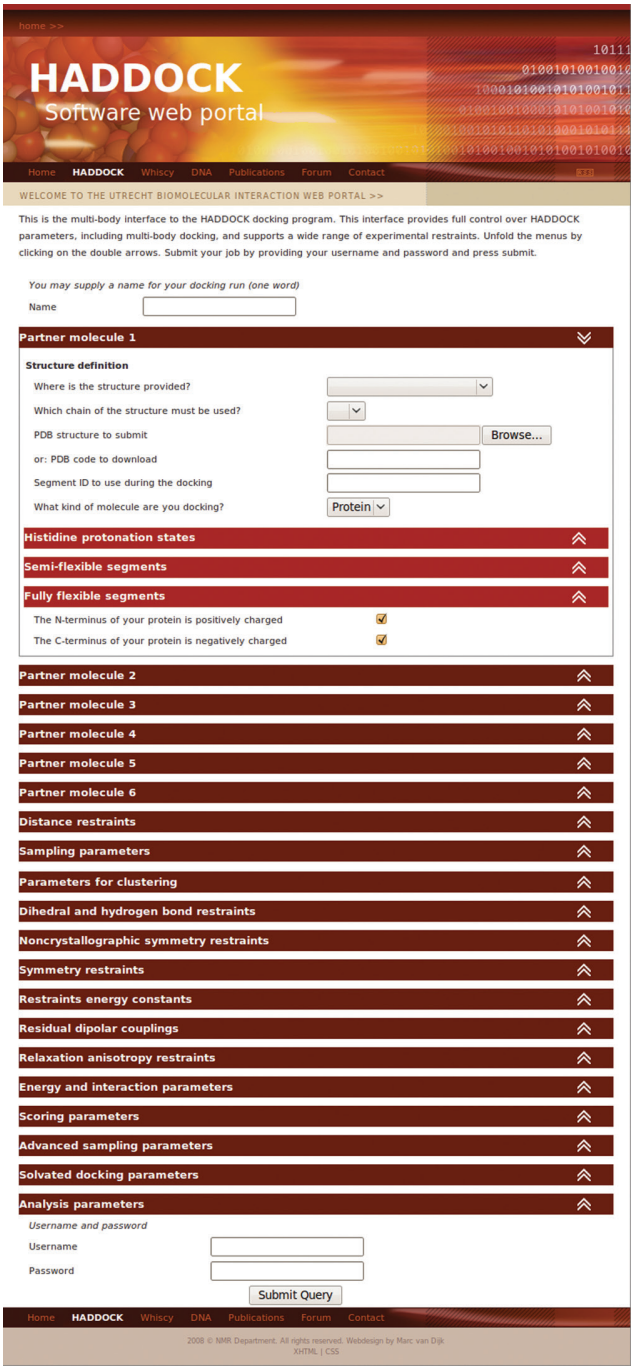



**Figure 2: View of the HADDOCK's multi-body web interface for data-driven docking.**

**[ http://haddock.chem.uu.nl/services/HADOCK ]**

The performance of our multibody docking approach was demonstrated for six complexes (**Table 3**) using a combination of experimental and/or bioinformatics predictions. For four of the complexes (1QU9, 1OUS, 1VIM, and 1VPN), active and passive residues were defined based on consensus bioinformatics interface predictions from CPORT (see **Materials and Methods**). For the other two complexes, a combination of experimental and predicted information was used. The list of *active* and *passive* residues for each complex is given in **Table 4**. Using that information, the HADDOCK multibody server produced and ranked at the top high quality models, demonstrating the excellent performance of our approach. Both the top ranked models and the top ranked clusters according to the HADDOCK score contained at least a medium quality (two-star) prediction (see **Table 5**).



Furthermore, analysis of the results showed that the imposed symmetries are fulfilled. In four cases, bound docking was performed, including a trimer (1QU9), two tetramers (1OUS and 1VIM), and a pentamer (1VPN). For each of them, the first ranked HADDOCK model corresponds to a high quality prediction. Considering the increased docking complexity due to the large interaction space to be sampled in the case of multicomponent systems, this demonstrates an outstanding performance. In the two unbound cases consisting of a CAPRI target and a protein-DNA complex, good results (two-star quality predictions) were also obtained (see **Table 5**). The performance of the protein-DNA docking (3CRO) and the ability of HADDOCK to catch the conformational changes in the DNA demonstrate that the excellent capabilities of HADDOCK are not limited to just protein-protein complexes. An overlay of the top predictions onto their respective reference crystal structures is shown in **Figure 3**.

**Table 4. List of active and passive residues used in HADDOCK in order to dock the various benchmark complexes.**

| PDB ID | Active Residues | Passive Residues |
|---|---|---|
| 1QU9[a] | 3,4,6,7,8,11-18,21,28-31,33,69,72,73,75,77,81,82,85,88,92,100-114,120,122,124 | 2,9,23,24,26,36,37,38,42,52,58,63,64,67,70,79,80,83,86,89,90,93,96,97,98,99,115,116,118,126-128 |
| 1URZ[b] | 5,8,9,10,11,13,54,71,73,75,76,78,79,87,93,98,110,118,193,196,219,222,244,248,251,267,269,270 | 4,7,12,15,21,22,24,26,28,34,36,56,57,64,66-70,72,77,81,83,86,92,94-96,107,108,120,131,150,152,154,192,194,195,216-218,224,243,246,250,253-263,266,271,272,273 |
| 1OUS[a] | 3,15,17,19,41,42,47-52,71,76-87,89,91,93,98,99,100-103,106,108,110,112-114 | 1,2,5,7,9,12,13,21,24,25,27,39,43,45,46,53,54,64,66,68,69,70,72,73-75,96,97 |

**Table 4. Continued.**

| | | | |
|---|---|---|---|
| 1VIM[a] | -2-10,16,41-47,50,51,54,55,57,63-73,138,140-142,144,145,147,150,151,154,155,158,159,162,163,176-185 | | 12-19,35,60-62,74,75,89,91,95,102,129,133-137,146,165-168,170-174 |
| 1VPN[a] | 32-38,52,71,74,75,78,79,107,111-119,123,127,130-137,139,142,152,160,162,225,228,229,239,240-245,250,252-260,264-269,274,275,288,289,291,296,299,300,303,314,316 | | 39-41,50,51,54,56,58,60,63-68,72,73,77,80,81,88,93,101,102,104-106,108-110,117,124,126,128,138,140,141,143-146,150,151,153-156,158,170,177,179,183,185,231-236,238,244,246-249,251,261,262,276,290,292-295,297,305,307,309,311,312 |
| 3CRO[c] | Protein: 29,31,32,42-44 | DNA: 4-7,13-18,22-25,32-36 | Protein: 9,18-20,27,28,30,34,36,37,40,41,45,46 |

The *active* and *passive residue* information is gathered via [a]CPORT, [b]CPORT and literature data, [c]conservation and experimental data (mutagenesis, ethylation interference).

**Table 5. Multi-body docking results obtained via using the multi-body interface of HADDOCK web-server[a].**

| | Quality/ Rank | Best structure i-RMSD / l-RMSD (Å) | Best cluster Quality/Rank | Best cluster i-RMSD / l-RMSD (Å) |
|---|---|---|---|---|
| 1QU9[b] | ★★★ / 1 | 0.8 / 0.7 | ★★★/1 | 0.8±0.1 / 0.7±0.1 |
| 1URZ[ub] | ★★/ 1 | 1.7 / 5.2 | ★★/1 | 1.8±0.1 / 5.3±0.1 |
| 1OUS[b] | ★★★/ 1 | 0.9 / 1.2 | ★★★/1 | 0.8±0.1 / 1.3±0.6 |
| 1VIM[b] | ★★★/ 1 | 1.0 / 1.2 | ★★★/1 | 1.2±0.2 / 1.3±0.2 |
| 1VPN[b] | ★★★/ 1 | 0.7 / 0.7 | ★★/1 | 4.1±0.1 / 4.0±0.1 |
| 3CRO[ub] | ★★/ 1 | 1.79 / 2.2 | ★★/1 | 2.12±0.3 / 2.8±0.6 |

[a]For the definition of i-RMSD and l-RMSD refer to the Methodology section, [b]Bound docking – the docking was performed with the separated monomers taken from the reference crystal structure, [c]Unbound docking – the docking was performed with the free form of the monomers (see **Materials and Methods** for details).

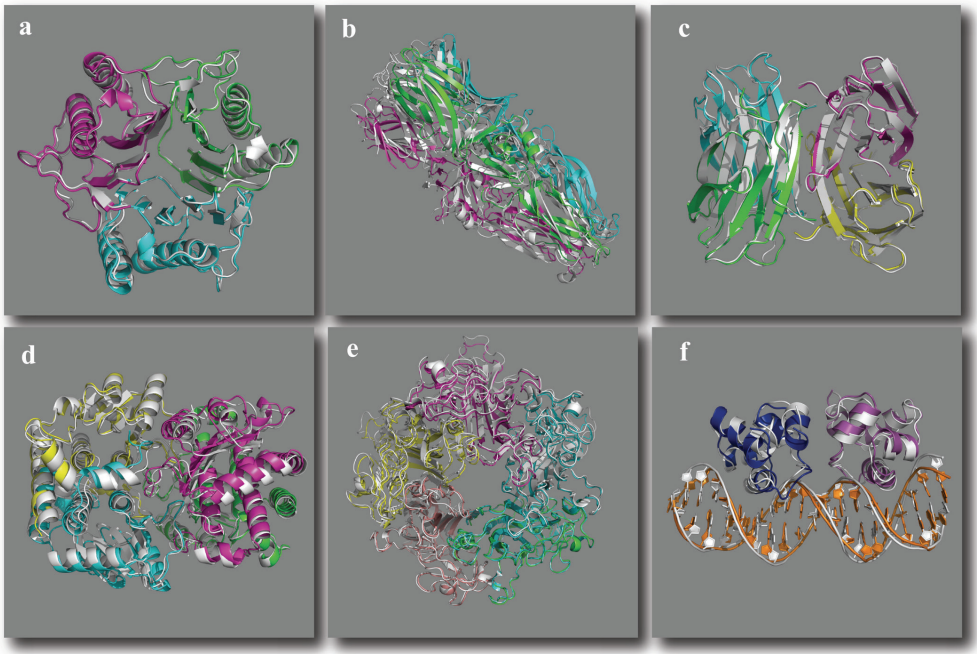



**Figure 3. View of the best HADDOCK solutions (having colored monomers) superimposed onto their respective crystal reference structures (shown in light grey). a.** 1QU9 **b.** 1URZ **c.** 1OUS **d.** 1VIM **e.** 1VPN and **f.** 3CRO. The figures were generated with Pymol (Delano Scientific LLC, http://www.pymol.org).

## 4. Discussion

### *4.1. Pushing Back Limits of Structural Prediction of Macromolecular Assemblies*

In the structural characterization of biomolecules, most of the developments on the modeling side are limited to rather "small" binary systems often only applicable to proteins. Just a few molecular docking programs can deal with multibody assemblies, and they are generally restricted to the prediction of symmetric homomeric complexes [32–37]. So far, HADDOCK [26, 28, 43] is the only molecular docking program that is able to perform simultaneous docking of multibody complexes up to six components. A multibody docking server was recently released on the HADDOCK web portal, allowing the users, through a user-friendly interface, to exploit the full range of experimental data supported by HADDOCK and to fully customize the docking process. The performance of this web server was evaluated in this study against a benchmark set of six multimeric complexes, including a protein-DNA complex. Cyclic or dihedral symmetries, which are present in the large majority of homomers [44], were defined combined with the interface information derived from experimental evidences and/or predictions made

by our consensus interface prediction program, CPORT. The results show that HADDOCK is able to generate native to near-native predictions for all cases with i-RMSD values for the best model ranging between 0.7 and 2.2 Å. Although we could produce excellent results even with the inclusion of bioinformatics predictions, which usually contain a considerable amount of false positives, one should always keep in mind that the information supplied to HADDOCK should be of high quality. This is because the complexity of the interaction space is larger in the case of multibody docking compared with the two-body docking.



The vast quantity of low-resolution experimental data that could further be used in HADDOCK paves the route for the prediction of large macromolecular complexes. By combining distance and interaction restraints from low-resolution methods with molecular docking, architectural or even atomic models might be generated. These restraints can be derived from a variety of experimental measurements including MS of intact complexes, chemical cross-linking, cryo-EM, SAXS, fluorescence resonance energy transfer, and analytical ultracentrifugation. One very recent addition to this series of biophysical tools is ion mobility separation (IM) coupled to MS [45]. IM is an established technique for studying shape and conformation of small molecules and individual proteins. When coupled with MS, mass and subunit composition of a protein complex can be determined simultaneously with its overall topology and shape [46, 47]. The cross-sections of amyloid oligomers formed in the early steps of amyloid fibril formation calculated by IM-MS [48] could be used as a restraint in data-driven docking to discriminate between quaternary topologies for a specific oligomeric state. This can be done for example by inferring a radius of gyration restraint from this cross-section measurement or by predicting the cross-sections from the docking models and using the experimental data as a filter.

### *4.2. Are We Ready to Predict Interactomes from Three-dimensional Structures of Biomolecules?*

In today's proteomics era, large-scale screening techniques are used to characterize protein-protein interactions (PPIs) in vivo [49, 50]. Despite the massive number of interactions detected by protein complex purification techniques using MS (originally either by high throughput MS protein complex identification [51] or by high throughput mass spectrometry protein complex identification coupled to tandem affinity purification [52], systematic yeast two-hybrid screening [53–55], complementary mapping techniques (e.g. protein fragment complementation assay [56], and in vitro proteome chip screening [57])), the interactome coverage remains low, roughly 50 and 10% for the yeast and human interactomes, respectively [58].

This becomes apparent in the rather limited overlap between various data sets obtained with different approaches [59]. This can be explained by a limit in proteome coverage (up to 70% for the best approaches) and by the inherent high fraction of false positives (previous estimations mention that more than half of all current high throughput data are spurious [59]). It also highlights the difficulties encountered by some methods for certain types of interactions, strengthening the complementarities between the different techniques. Finally, proteomics data sets derived to map PPIs, even when a similar detection method is used [7, 60], have a limited overlap (only 18%) [61]. Computational methods to predict protein assemblies could in principle play a complementary role in the study of interactomes, providing additional insights with leverage of the structural models. But can present scoring functions used in protein-protein docking methods characterize the binding affinity of a macromolecular complex, a requisite to predict interactomes? To answer this question, we have tested nine of the currently best performing scoring functions against a large set of high quality binding affinity data derived from the literature [62]. The results (data not shown) reveal that scoring is orthogonal to binding affinity prediction (the highest calculated $r^2$ was 0.09!) even though scoring functions are successfully being used in discriminating native structure from decoys. Hence, even if structural modeling tools and molecular docking approaches can significantly improve the selection accuracy of PPI networks [63], these computational methods need to be optimized for both purposes, e.g. annotation and prediction of PPIs.



### *4.3. Need for Combining Experimental Information and Modeling*

By combining a variety of experimental approaches, one can easily increase our knowledge about biologically relevant interaction [64]. The experimental information can guide large scale docking studies to upgrade the information contained in interactome maps by adding the three-dimensional structural dimension to the PPIs. Moving toward systems biology, computational methods could aim at predicting how the proteome is wired and how dynamic changes in the interactome occur in response to different environmental factors. In that regard, mass spectrometry techniques that determine the composition and stoichiometry of macromolecular complexes will be of indispensable value.

But how far are we from a high throughput method to screen for protein complex structures? Recently, we have linked HADDOCK to MTMDAT, an automated software for the analysis of mass spectrometry data [65], creating effectively a pipeline for high throughput, MS-based structural modeling of complexes. This pipeline allows feeding automatically into HADDOCK the interface

information identified by MS from digestion experiments. This is only one example of how experiments and modeling can be coupled, and we expect that many other related applications will be developed in the future to open the route to large-scale annotation of interactomes.

**Acknowledgments—** We thank Dr. Marc van Dijk (Utrecht University) for providing the data for the protein-DNA complex discussed in this work.

## References

[1] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860-921.
[2] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., et al. (2001) The sequence of the human genome. Science 291, 1304-1351.
[3] Xie, L., and Bourne, P. E. (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. PLoS Comput. Biol. 1, e31.
[4] Alber, F., Forster, F., Korkin, D., Topf, M., and Sali, A. (2008) Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem. 77, 443-477.
[5] Mueller, M., Jenni, S., and Ban, N. (2007) Strategies for crystallization and structure determination of very large macromolecular assemblies. Curr. Opin. Struct. Biol. 17, 572-579.
[6] Collins, S. R., Kemmeren, P., Zhao, X. C., Greenblatt, J. F., Spencer, F., Holstege, F. C., Weissman, J. S., and Krogan, N. J. (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol. Cell. Proteomics 6, 439-450.
[7] Krogan, N. J., Cagney, G., Yu, H. Y., Zhong, G. Q., Guo, X. H., Ignatchenko, A., Li, J., Pu, S. Y., Datta, N., et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440, 637-643.
[8] van Dijk, A. D., Kaptein, R., Boelens, R., and Bonvin, A. M. (2006) Combining NMR relaxation with chemical shift perturbation data to drive protein-protein docking. J. Biomol. NMR 34, 237-244.
[9] van Dijk, A. D., Fushman, D., and Bonvin, A. M. (2005) Various strategies of using residual dipolar couplings in NMR-driven protein docking: application to Lys48-linked di-ubiquitin and validation against 15N-relaxation data. Proteins 60, 367-381.
[10] Cunningham, B. C., Jhurani, P., Ng, P., and Wells, J. A. (1989) Receptor and antibody epitopes in human growth hormone identified by homolog-scanning mutagenesis. Science 243, 1330-1336.
[11] Peters, K., and Richards, F. M. (1977) Chemical cross-linking: reagents and problems in studies of membrane structure. Annu. Rev. Biochem. 46, 523-551.
[12] Janin, J., Henrick, K., Moult, J., Ten Eyck, L., Sternberg, M. J. E., Vajda, S., Vasker, I., and Wodak, S. J. (2003) CAPRI: A Critical Assessment of PRedicted Interactions. Proteins 52, 2-9.
13] Mendez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003) Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 52, 51-67.
[14] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992) Molecular-Surface Recognition - Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. Proc. Natl. Acad. Sci. U. S. A. 89, 2195-2199.
[15] Ben-Zeev, E., and Eisenstein, M. (2003) Weighted geometric docking: Incorporating external information in the rotation-translation scan. Proteins 52, 24-27.
[16] Zacharias, M. (2003) Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci. 12, 1271-1282.
[17] Zacharias, M. (2004) Rapid protein-ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: binding of FK506 to FKBP. Proteins: Struct., Funct., Bioinf. 54, 759-767.
[18] Chen, R., Li, L., and Weng, Z. (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52, 80-87.
[19] Pierce, B., and Weng, Z. (2007) ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 67, 1078-1086.

[20] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 33, W363-W367.
[21] Schneidman-Duhovny, D., Inbar, Y., Polak, V., Shatsky, M., Halperin, I., Benyamini, H., Barzilai, A., Dror, O., Haspel, N., et al. (2003) Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. Proteins: Struct., Funct., Bioinf. 52, 107-112.
[22] Grosdidier, S., Pons, C., Solernou, A., and Fernandez-Recio, J. (2007) Prediction and scoring of docking poses with pyDock. Proteins 69, 852-858.
[23] Man-Kuang Cheng, T., Blundell, T. L., and Fernandez-Recio, J. (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 68, 503-515.
[24] Gerega, S. K., and Downard, K. M. (2006) PROXIMO--a new docking algorithm to model protein complexes using data from radical probe mass spectrometry (RP-MS). Bioinformatics 22, 1702-1709.
[25] Lasker, K., Topf, M., Sali, A., and Wolfson, H. J. (2009) Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. J. Mol. Biol. 388, 180-194.
[26] 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: Struct., Funct., Bioinfo. 69, 726-733.
[27] De Vries, S. J., Van Dijk, M., and Bonvin, A. M. (in Press) The HADDOCK web server for data-driven biomolecular docking. Nat. Protoc.
[28] 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.
[29] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235-242.
[30] Ratnaparkhi, G. S., and Varadarajan, R. (2001) Osmolytes stabilize ribonuclease S by stabilizing its fragments S protein and S peptide to compact folding-competent states. J. Biol. Chem. 276, 28789-28798.
[31] Berchanski, A., and Eisenstein, M. (2003) Construction of molecular assemblies via docking: modeling of tetramers with D2 symmetry. Proteins 53, 817-829.
[32] Berchanski, A., Segal, D., and Eisenstein, M. (2005) Modeling oligomers with Cn or Dn symmetry: application to CAPRI target 10. Proteins 60, 202-206.
[33] Comeau, S. R., and Camacho, C. J. (2005) Predicting oligomeric assemblies: N-mers a primer. J. Struct. Biol. 150, 233-244.
[34] Andre, I., Bradley, P., Wang, C., and Baker, D. (2007) Prediction of the structure of symmetrical protein assemblies. Proc. Natl. Acad. Sci. U. S. A. 104, 17656-17661.
[35] Pierce, B., Tong, W. W., and Weng, Z. P. (2005) M-ZDOCK: a grid-based approach for C-n symmetric multimer docking. Bioinformatics 21, 1472-1478.
[36] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005) Geometry-based flexible and symmetric protein docking. Proteins: Struct., Funct., Bioinf. 60, 224-231.
[37] Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H. J. (2005) Prediction of multimolecular assemblies by multiple docking. J. Mol. Biol. 349, 435-447.
[38] Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse_Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., et al. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr., Sect. D: Biol. Crystallogr. 54 ( Pt 5), 905-921.

[39] Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2004) Identification of protein-protein interaction sites from docking energy landscapes. J. Mol. Biol. 335, 843-865.
[40] Brunger, A. T. (1992) X-PLOR. A System for X-Ray Crystallography and NMR, Yale University Press, New Haven, CT.
[41] Nilges, M. (1993) A Calculation Strategy for the Structure Determination of Symmetrical Dimers by H-1-Nmr. Proteins 17, 297-309.
[42] O'Donoghue, S. I., and Nilges, M. (1999) Structure Computation and Dyanmics in Protein NMR, Kluwer Academic/Plenum Publishers, New York.
[43] van Dijk, A. D. J., de Vries, S. J., Dominguez, C., Chen, H., Zhou, H. X., and Bonvin, A. M. J. J. (2005) Data-driven docking: HADDOCK's adventures in CAPRI. Proteins: Struct., Funct., Bioinf. 60, 232-238.
[44] Levy, E. D., Erba, E. B., Robinson, C. V., and Teichmann, S. A. (2008) Assembly reflects evolution of protein complexes. Nature 453, 1262-1265.
[45] Ruotolo, B. T., Giles, K., Campuzano, I., Sandercock, A. M., Bateman, R. H., and Robinson, C. V. (2005) Evidence for macromolecular protein rings in the absence of bulk water. Science 310, 1658-1661.
[46] Ruotolo, B. T., Benesch, J. L. P., Sandercock, A. M., Hyung, S. J., and Robinson, C. V. (2008) Ion mobility-mass spectrometry analysis of large protein complexes. Nat. Protoc. 3, 1139-1152.
[47] Smith, D. P., Knapman, T. W., Campuzano, I., Malham, R. W., Berryman, J. T., Radford, S. E., and Ashcroft, A. E. (2009) Deciphering drift time measurements from travelling wave ion mobility spectrometry-mass spectrometry studies. European Journal of Mass Spectrometry 15, 113-130.
[48] Bernstein, S. L., Dupuis, N. F., Lazo, N. D., Wyttenbach, T., Condron, M. M., Bitan, G., Teplow, D. B., Shea, J. E., Ruotolo, B. T., et al. (2009) Amyloid-beta protein oligomerization and the importance of tetramers and dodecamers in the aetiology of Alzheimer's disease. Nat. Chem. 1, 326-331.
[49] Walhout, A. J. M., and Vidal, M. (2001) Protein interaction maps for model organisms. Nat. Rev. Mol. Cell Biol. 2, 55-62.
[50] Auerbach, D., Fetchko, M., and Stagljar, I. (2003) Proteomic approaches for generating comprehensive protein interaction maps. TARGETS 2, 85-92.
[51] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415, 180-183.
[52] Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141-147.
[53] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623-627.
[54] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. U. S. A. 98, 4569-4574.
[55] Yu, H. Y., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322, 104-110.
[56] Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M. S., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., et al. (2008) An in vivo map of the yeast protein interactome. Science 320, 1465-1470.

[57] Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., et al. (2001) Global analysis of protein activities using proteome chips. Science 293, 2101-2105.
[58] Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006) How complete are current yeast and human protein-interaction networks? Genome Biology 7, 120.1-120.9.
[59] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399-403.
[60] Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631-636.
[61] Goll, J., and Uetz, P. (2006) The elusive yeast interactome. Genome Biol. 7, 223.1-223.6.
[62] Kastritis, P. L., and Bonvin, A. M. J. J. (in Press) Are scoring functions in protein protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J. Proteome Res.
[63] Fernandez-Ballester, G., and Serrano, L. (2006) Prediction of protein-protein interaction [based on structure. Methods in Molecular Biology 340, 207-234.
[64] Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295, 321-324.
[65] Hennig, J., Hennig, K. D. M., and Sunnerhagen, M. (2008) MTMDAT: Automated analysis and visualization of mass spectrometry data for tertiary and quaternary structure probing of proteins. Bioinformatics 24, 1310-1312.
[66] Volz, K. (1999) A test case for structure-based functional assignment: The 1.2 angstrom crystal structure of the yjgF gene product from Escherichia coli. Protein Sci. 8, 2428-2437.
[67] Bressanelli, S., Stiasny, K., Allison, S. L., Stura, E. A., Duquerroy, S., Lescar, J., Heinz, F. X., and Rey, F. A. (2004) Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. EMBO J. 23, 728-738.
[68] Loris, R., Tielker, D., Jaeger, K. E., and Wyns, L. (2003) Structural basis of carbohydrate recognition by the lectin LecB from Pseudomonas aeruginosa. J. Mol. Biol. 331, 861-870.
[69] Badger, J., Sauder, J. M., Adams, J. M., Antonysamy, S., Bain, K., Bergseid, M. G., Buchanan, S. G., Buchanan, M. D., Batiyenko, Y., et al. (2005) Structural analysis of a set of proteins resulting from a bacterial genomics project. Proteins: Struct., Funct., Bioinf. 60, 787-796.
[70] Stehle, T., and Harrison, S. C. (1997) High-resolution structure of a polyomavirus VP1-oligosaccharide complex: implications for assembly and receptor binding. Embo J. 16, 5139-5148.
[71] Mondragon, A., and Harrison, S. C. (1991) The Phage-434 Cro/Or1 Complex at 2.5Å Resolution. J. Mol. Biol. 219, 321-334.

Chapter

3

# A Multidomain Flexible Docking Approach

## to Deal with Large Conformational Changes in the Modeling of Biomolecular Complexes

Based on research article:
**Ezgi Karaca, Alexandre M.J.J. Bonvin,** A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. *Structure*, 2011, 19(4):555-65.

**Abstract**

Binding-induced backbone and large-scale conformational changes represent one of the major challenges in the modeling of biomolecular complexes by docking. To address this challenge, we have developed a *Flexible Multidomain Docking* protocol that follows a "divide-and-conquer" approach to model both large-scale domain motions and small- to medium-scale interfacial rearrangements: the flexible binding partner is treated as an assembly of sub-parts/domains that are docked simultaneously making use of HADDOCK's multibody docking ability. For this, the flexible molecules are cut at hinge regions predicted using an elastic network model. The performance of this approach is demonstrated on a benchmark covering an unprecedented range of conformational changes of 1.5 to 19.5Å. We show from a statistical survey of known complexes that the cumulative sum of eigenvalues obtained from the elastic network has some predictive power to indicate the extent of the conformational change to be expected.

## 1. Introduction

Proteins play a vital role in all kinds of biological processes through their complex interactions with other biomolecules and small ligands. Revealing all the functional steps in the lifetime of a protein requires 3D atomic-level information about the complexes it forms. This information can be obtained from classical experimental techniques such as X-ray crystallography and NMR, although, considering the enormous number of complexes, the need for accurate and complementary computational methods like docking is evident [1–6]. Docking is defined as the modeling of the structure of a complex (the bound conformation) starting from the free forms (unbound conformations) of the interaction partners. This becomes an extremely challenging task when the conformation of a protein changes significantly upon complex formation [1,2,6,7].



Various mechanisms have been proposed over the years to describe the binding process and its associated conformational changes. Fischer suggested that the active sites of the interacting molecules are complementary to each other, obeying a *lock-and-key* mechanism [8]. Koshland extended this model by proposing that the binding follows an *induced fit* rule, where the molecules adopt different conformations in order to fit into the active site of their interaction partner [9]. Later on, Kumar *et al.* reevaluated this issue from a statistical mechanics point of view and proposed the *conformational selection* model [10]: The native state of a protein exists in an ensemble of conformations sampling various states and environmental effects shift the equilibrium towards the bound conformation. Subsequent to this hypothesis, Grunberg *et al*. [11] provided a *consensus model*, treating the concept of recognition and binding as two different processes with conformational selection at the basis of the recognition process; starting with the formation of an encounter complex followed by an induced fit mechanism leading to the final complex.

Conformational changes occurring upon binding can be classified into four categories: local rearrangements, collective global motions, a mixture of the two and binding-induced folding events. Local changes cover loop rearrangements [7], secondary structure element alterations and motions etc. Global changes refer to large scale domain motions like hinge and shear [7,12,13]. Binding-induced folding events are observed in specific recognition when the folding of one of (partially) unfolded partner occurs only upon binding [1,6,14]. Various strategies have been followed to model these different types of conformational changes [1–3,6,7]. For the *lock-and-key* cases, namely when shape complementarily drives the interaction, rigid body algorithms (e.g. geometric hashing [15] and Fast Fourier Transform [16]) provide fast and accurate solutions. Small- to medium-scale backbone conformational changes (in the range of 1-2Å) can be modeled through algorithms

that address the *induced fit* mechanism (mostly based on Molecular Dynamics and Monte Carlo simulations [17,18]). For some types of conformational changes, e.g. loop rearrangements or allosteric effects, this may not provide satisfactory results. Under such conditions, docking can be started from ensembles of structures (coming from NMR, Molecular Dynamics or Monte Carlo Simulations, Normal Mode Analysis, graph theoretic approaches, etc.), following thus the conformational selection model [17,19,20]. As an alternative method, a *multi-copy mean-field* approach [21,22] can also be used. The final class of algorithms is based on the consensus binding model introduced by Grunberg *et al.* [11]: structures are docked starting from an ensemble of conformations and then refined allowing for small conformational changes to take place [19]. This class of algorithms can be useful if different types of conformational changes are observed at the same time. Modeling of the last class of conformational changes, namely folding upon binding should be considered separately from regular docking as it requires incorporation of protein structure prediction within the context of docking. As a first step to deal with this problem, Baker's group recently developed a *fold-and-dock* protocol within Rosetta for modeling symmetric homo-oligomers [23].



All of the above mentioned algorithms work reasonably well for small- to medium-scale local conformational changes; but they usually fail to model large loop rearrangements, secondary structure changes and collective domain motions where the backbone RMSD of a protein changes by more than 2-3Å [1,2,24]. This is mainly due to the complexity of the conformational space to be sampled and the difficulty of predicting *a priori* such type of conformational changes [1,2,24]. Hybrid methodologies that combine different approaches are thus needed in order to address this problem. Until now five hybrid methods have been proposed to model large-scale conformational changes: ATTRACT [22,25], SwarmDock [26,27], multibody multistage docking procedure of MolFit [28], FlexDock [29] and fold-tree representation of Rosetta [30]. ATTRACT can incorporate soft harmonic low frequency modes into the docking procedure and provide fast relaxation/adaptation of the structure on a global scale [25]. SwarmDock also uses normal modes in the docking procedure but, in contrast to ATTRACT, it takes a linear combination of both low and high frequency normal modes into account for addressing high frequency thermal motions occurring at the interface [27]. MolFit treats the individual domains of the molecules as soft rigid objects and then docks them in a sequential multi-stage two-body docking protocol [28]. The method of FlexDock exploits a similar methodology, it dissects the flexible protein into rigid domains and performs a pairwise docking of the separate domains using PatchDock [31] followed by assembly of the resulting models. In Rosetta, the protein is represented as a fold-tree, which provides a mean for defining flexible regions between centers of the rigid

molecules and thus allowing domains to move with respect to each other during rigid-body optimization [30]. Next to these docking approaches, there is a recently published flexible refinement server, FiberDock constructed to deal with large conformational changes. FiberDock deforms the backbone in the direction of the selected normal modes [32]. The normal mode selection is based on the correlation of that particular mode with the van der Waals forces. All these approaches are quite promising, but they also have some limitations, especially in treating interfacial induced backbone and side-chains conformational changes at the same time.

Our in house docking software HADDOCK has already proven its ability to deal with small- to medium-scale induced conformational changes by combining docking from ensembles of starting structures with flexible refinement of both side-chains and backbone [17,33,34]. Moreover we have recently demonstrated assemblies [35]. Here we combine all these aspects and propose a straightforward and easy-to-apply docking protocol that can deal with large conformational changes while accounting for local changes at the same time. Following a "divide-and-conquer" approach, our *Flexible Multidomain Docking* (FMD) protocol within HADDOCK partitions the flexible molecule into rigid bodies with connectivity restraints between them and performs a simultaneous multibody docking of all components [35]. The proteins are dissected into domains by cutting them at hinge regions predicted from an Elastic Network Model [36]. This allows modeling of global scale changes at the rigid body docking stage. The resulting models are subsequently subjected to a flexible refinement involving both side-chains and backbone motions to deal with small- to medium-scale induced conformational changes. Our FMD protocol was tested against a benchmark of eleven protein-protein complexes that are experiencing domain motions, spanning a range of conformational changes from 1.5Å to as much as 19.5Å. This new FMD protocol is shown to be an excellent approach to model conformational changes as large as 19.5Å.

## 2. Experimental Procedures

### *2.1. Determination of the Hinge Regions*

The hinge prediction server HingeProt [36] was used to define the hinge regions of the flexible monomers. HingeProt annotates rigid parts and possible hinge regions of the supplied protein based on two elastic network models: GNM [37] and Anisotropic Network Model [38]. In this work, two different outputs provided by the HingeProt server are used: the predicted hinge regions and the eigenvalues obtained by the decomposition of the connectivity (Kirchhoff) matrix [37]. The hinge predictions obtained from HingeProt are filtered in order to preserve the secondary

structure integrity. Insignificant and/or structurally unreliable predictions are eliminated. The detailed procedure is described in the **Results** section.

### *2.2. Docking Protocol of HADDOCK*

Both two-body docking and FMD were performed with HADDOCK 2.1 [17,34,35]. The interface information used to construct the AIRs was extracted from the crystal structure of the complexes. We assumed to have an ideal definition of the interacting surfaces in order to concentrate on our ability to deal with large conformational changes. The AIR definitions are provided in **Table S3**. For the runs with bioinformatics predictions, we used CPORT [39]. In the FMD runs, center-of-mass restraints between the domains were turned on to ensure compactness of the solutions. Center-of-mass restraints are defined as a distance restraint between the geometric average positions of all Ca atoms within each molecule. The distance is automatically defined based on the dimension of the molecules or domains. The number of structures was increased to 5000, 400 and 400 for it0, it1 and water respectively. Random removal of AIRs was turned off since ideal restraints were used. Other parameters were left to their default values. Scoring and clustering were performed according to standard HADDOCK procedures [17,34].



### *2.3. Assessment of the structure quality*

The docking models were evaluated according to CAPRI criteria [40]:
- Acceptable prediction (one star): i-RMSD≤4Å or l-RMSD≤10Å and Fnat≥ 0.1
- Good prediction (two stars): i-RMSD≤2Å or l-RMSD≤5Å and Fnat≥0.3
- High quality prediction (three stars): i-RMSD≤1Å or l-RMSD≤1Å and nat≥0.5

i-RMSD refers to the interface RMSD, l-RMSD to the ligand RMSD, calculated over the backbone atoms of the ligand (rigid component) after fitting on the receptor, and Fnat to the fraction of native contacts. Next to Fnat we also provided Fnonnat, which is the fraction of incorrectly predicted contacts given the modeled complex. A cluster was considered of one-, two- or three-star quality if at least one of its top four members was of the corresponding quality.

## 3. Results

### *3.1. Benchmark Compilation*

The FMD benchmark was compiled according to three major criteria: (i) One of the partners should only undergo a small conformational change, of less than 2.0 Å, in order to decrease the level of complexity, (ii) the other partner should

experience conformational change emanating from hinge motions (ideally from one hinge position) and the hinge motion should be functionally involved in the binding process, (iii) both the bound and unbound conformations of the partners should be available.

Eleven protein-protein complexes were found that fit all the above-mentioned requirements, except 1NPE, a former Critical Assessment of PRediction of Interactions (CAPRI [40]) target, with only the bound conformation of the ligand available (**Table 1**). The other complexes were taken from the Protein-Protein Docking Benchmark 4.0 [73], eight of which belonging to the difficult category. A vast range of conformational changes is covered, from 1.5Å to 19.5Å, with 1IRA, 1H1V, 1Y64 1F6M, and 1FAK being particularly challenging (**Table 1**, **Figure 1**).



**Table 1. Selected complexes for the Flexible Multidomain Docking Benchmark.**

| | | | Backbone Conformational Change Range (Å) | |
|---|---|---|---|---|
| **Complex ID** | **Receptor ID[a]** | **Ligand ID[a]** | **Receptor** | **Ligand** |
| 1IRA[uu] [41] | 1G0Y_R [42] | 1ILR_1 [43] | 19.5 | 0.7 |
| 1H1V[uu] [44] | 1D0N_B [45] | 1IJJ_B [46] | 13.9 | 1.6 |
| 1Y64[uu] [47] | 1UX5_A [48] | 2FXU_A [49] | 10.3 | 1.1 |
| 1F6M[uu] [50] | 1CL0_A [51] | 2TIR_A [52] | 7.3 | 0.9 |
| 1FAK[uu] [53] | 1QFK_HL [54] | 1TFH_B [55] | 6.0 | 1.0 |
| 1ZLI[uu] [56] | 2JTO_A [57] | 1KWM_A [58] | 3.8 | 0.6 |
| 1E4K[uu] [59] | 2DTQ_AB [60] | 1FNL_A [61] | 2.9 | 1.7 |
| 1IBR[u,] [62] | 1F59_A [63] | 1QG4_A [64] | 2.9 | 1.1 |
| 1KKL[uu] [65] | 1JB1_AB [66] | 2HPR [67] | 2.6 | 0.5 |
| 1NPE[ub] [68] | 1KLO_A [69] | 1NPE_A [68] | 1.8 | - |
| 1DFJ[uu] [70] | 2BNH_A [71] | 9RSA_B [72] | 1.5 | 0.7 |

[uu]Docking was performed starting from the unbound conformations of both receptor and ligand, [ub]For this particular protein, as the unbound conformation of the ligand was not available, the docking was performed from the unbound conformation of the receptor and the bound conformation of the ligand, [a]The PDB and chain ID's are indicated as PDB ID_chain ID.

### *3.2. The Workflow of Flexible Multidomain Docking*

#### *Dissecting proteins into domains*

Hinges were predicted using the HingeProt server [36]. They were then filtered to guarantee the structural integrity and dissect the flexible molecule in as few components as possible in order to maintain the compactness of the expected solution. This filtering is based on the domain motions study of Hayward in which

hinges were found to be more likely at the termini of α-helices, β-sheets and in loops that connect domains, less likely in the center of interdomain α-helices and not likely within β-sheets [75]. In addition, we also discarded hinge predictions close to N- and C-terminal residues, as it is highly improbable that such residues can have a hinge-like behavior. Subsequently, the monomer was cut at the first peptide bond following the predicted hinge residue. Short explanations for the basis of the hinge selection for each protein can be found in **Table 2**.

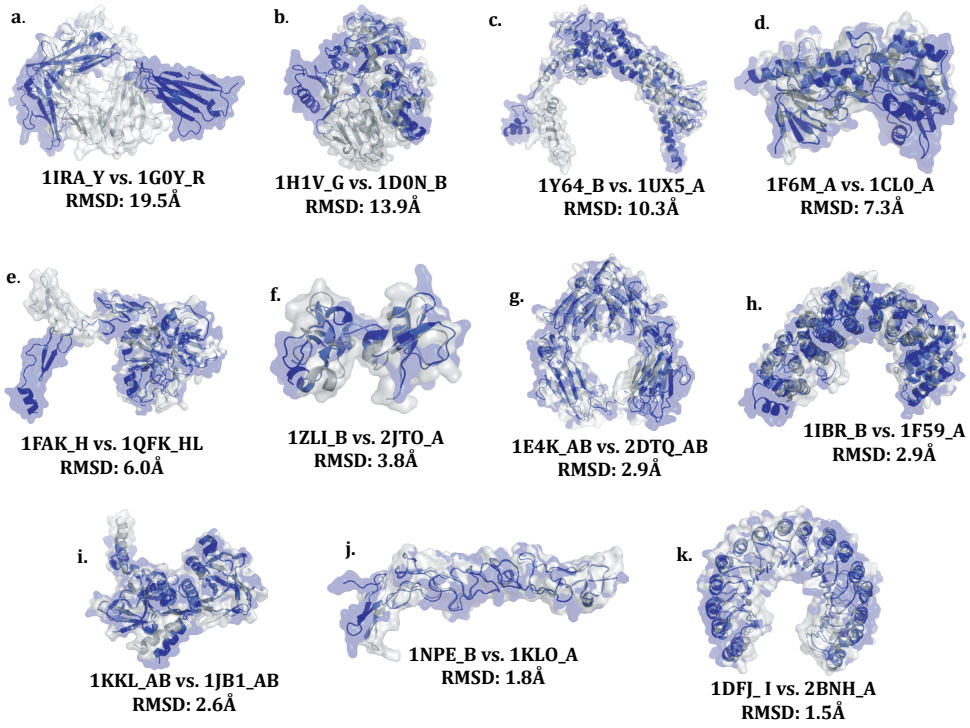



**Figure 1: Comparison of the unbound (grey) and bound (blue) superimposed conformations of the receptors in our benchmark. (a.** 1IRA **b.** 1H1V **c.** 1Y64 **d.** 1F6M **e.** 1FAK **f.** 1ZLI **g.** 1IBR **h.** 1E4K **i.** 1KKL **j.** 1NPE **k.** 1DFJ). For each case, the positional backbone RMSD between the two forms is indicated. The figures were generated with Pymol [74].

### *Defining ambiguous interaction and connectivity restraints*

The Ambiguous Interaction Restraints (AIRs) were generated via a new HADDOCK server interface especially built to allow fine-tuning of restraints in the case of multibody docking (http://haddock.chem.uu.nl/services/GenTBL/) [34]: (i) Restraints for the interface of the rigid molecule were kept ambiguous for the distinct interfaces of the separated flexible molecule, (ii) No AIRs were defined between the

separated domains of the flexible molecule (see **Figure 2**). The AIRs were defined based on interface residues identified from the crystal structure of the complexes. By choosing an ideal definition of the interface (but not of the contacts made) we focus on the problem of dealing with conformational changes. This represents thus a best-case scenario.

To maintain the connectivity between the separated chains, two connectivity restraint files consisting of an unambiguous distance restraint between the C- and N-termini of the separated domains were prepared. The first file was used during the rigid body energy minimization (*it0*) and semi-flexible refinement in torsion angle space (*it1*), to enforce a chain separation of no more than 10Å (upper distance restraint). The second one was imposed during the final refinement in explicit solvent (*water*) in order to restore the connectivity (to the real peptide distance, 1.3Å). As the N- and C-termini of the cut domains are created artificially, they were kept uncharged. Three residues on both sides of the separated domains were defined as fully flexible to allow for more flexibility of the linker region. All other HADDOCK parameters were left to their default values. The summary of the workflow is illustrated in **Figure 3**.

**Table 2. List of selected hinges and the basis of the selection.**

| Complex ID | Receptor ID | Hinge Predictions[a] (Selected hinge in bold) | Basis of the Selection |
|---|---|---|---|
| 1IRA_Y | 1G0Y_R | 13,98,**203**,307 | 13 and 307 are at the N- and C-termini, respectively. 98 and 203 are on a linker, but 203 is within a 10 residue longer linker than 98. |
| 1H1V_G | 1D0N_B | 423,446,500,534,**631** | Only 534 and 631 are on a linker, where 631 is more flexible compared to 534 according to experimentally determined B-Factors. |
| 1Y64_B | 1UX5_A | 1427,**1403**,1544,1647,1699 | 1403 is in a linker while the others are in α-helices. |
| 1F6M_A | 1CL0_A | 11,79,110,**115**,148,**245,**283,303 | This protein is cut at two places. 115 and 245 are selected as probable hinges, as they lie on a linker that connects two large domains. |
| 1FAK_H | 1QFK_L | 83,**89**,130 | 130 is a C-terminal residue. Both 83 and 89 are in the same linker; 89 was selected arbitrarily. |
| 1ZLI_B | 2JTO_A | **37** | The only hinge prediction is selected. |
| 1E4K_AB | 2DTQ_AB | 338,444(in chain A) **340**(in chain B) | The receptor of this complex is a symmetrical homodimer (composed of chain A and B). Except 444, which is at the C-terminus, both predictions point out the same regions in different chains, thus just one monomer is cut at 340. |
| 1IBR_B | 1F59_A | 116,**237**,322 | All predictions are in the middle of a-helices. Therefore 237 is selected, as it corresponds to the center of mass of the receptor. |
| 1KKL | 1JB1_ABC | 156,201,218,225,262,264,281,285,286(2 times),**291**,307 | The hinge prediction is run on chains A, AB and ABC. The receptor is a symmetric trimer, thus all the predictions are considered. The most frequently predicted hinge regions are around the 289-293 linker, in which 291 is located. |
| 1NPE_B | 1KLO_A | **43**,89,122 | In the literature it was already stated that this protein is composed of three modules, containing 4 disulphide bridges. Residue 122 forms a disulphide bridge. Residue 43 separates the first module, while residue 89 is in the middle of the second module [69]. |
| 1DFJ_I | 2BNH_A | **117**,229,326 | Residue 326 is in the middle of an α-helix, residue 229 is completely in the center of a barrel formed by α-helices and β-sheets and residue 117 is on a linker. |

[a]The hinge predictions were obtained with the HingeProt server [36].

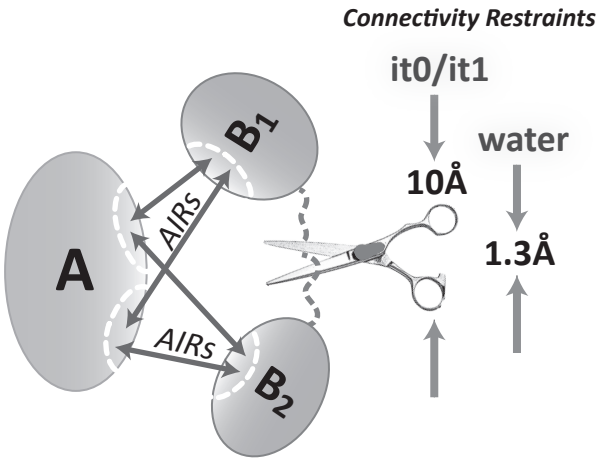



**Figure 2: Schematic representation of the restraints used in docking (ambiguous interaction and connectivity restraints) and of the dissection of the flexible partner into sub-domains.**

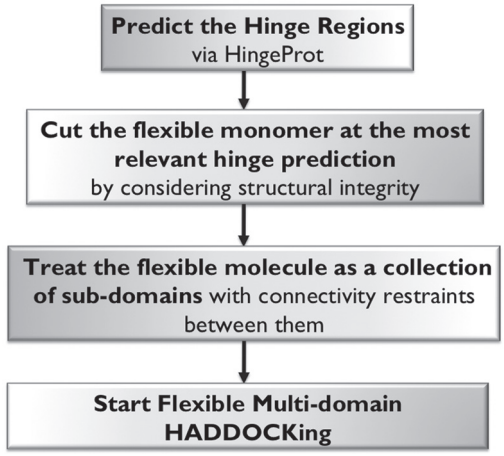



**Figure 3: Workflow of Flexible Multidomain Docking in HADDOCK**

### *3.3. Performance of Flexible Multidomain Docking*

Both standard semi-flexible two-body docking and the new FMD protocol were applied to the benchmark for comparison. Two-body docking could generate good solutions (two stars) for 1DFJ and 1NPE and acceptable ones (one star) for 1E4K and 1KKL. The best solutions were ranked at the top for the latter three (**Table 3**). For the other cases, two-body docking failed completely in generating any acceptable solutions. When the new FMD protocol was applied, acceptable or better solutions were obtained for each benchmark case (**Table 3**). The improvement is impressive particularly for the seven cases (1IRA, 1H1V, 1Y64, 1F6M, 1FAK, 1ZLI, 1IBR) for which two-body docking was failing: the best l-RMSD values of the top four challenging cases decreased from 38.6Å to 7.5Å for 1IRA, from 31.1Å to 9.6Å for 1H1V, from 3.2Å to 9.5Å for 1Y64 and from 39.4Å to 8.7Å

for 1F6M (Note also the very high fraction of native contacts (**Table 3**)).

**Table 3. Comparison of two-body and Flexible Multidomain Docking results.**[a]

| Flexible Multidomain Docking | | | | | |
|---|---|---|---|---|---|
| **PDB ID** | **Quality / Rank**[b] | **i-RMSD (Å)** | **l-RMSD (Å)** | **$F_{nat}$** | **$F_{nonnat}$** |
| 1IRA | ★ / 1 | 3.9 | 7.5 | 0.55 | 0.61 |
| 1H1V | ★ / 11 | 4.6 | 9.6 | 0.49 | 0.67 |
| 1Y64 | ★ / 5 | 3.9 | 9.5 | 0.48 | 0.72 |
| 1F6M | ★ / 1 | 3.5 | 8.7 | 0.69 | 0.58 |
| 1FAK | ★★ / 37 | 2.8 | 4.2 | 0.55 | 0.54 |
| 1ZLI | ★★ / 1 | 2.1 | 3.5 | 0.74 | 0.47 |
| 1E4K | ★★ / 5 | 2.3 | 4.0 | 0.70 | 0.47 |
| 1IBR | ★★ / 1 | 2.3 | 4.0 | 0.63 | 0.57 |
| 1KKL | ★★ / 1 | 2.2 | 4.9 | 0.67 | 0.59 |
| 1NPE | ★★ / 16 | 1.2 | 3.2 | 0.95 | 0.36 |
| 1DFJ | ★★ / 5 | 2.0 | 7.1 | 0.68 | 0.56 |

| 2-Body Docking[c] | | | | | |
|---|---|---|---|---|---|
| **PDB ID** | **Quality / Rank** | **i-RMSD (Å)** | **l-RMSD (Å)** | **$F_{nat}$** | **$F_{nonnat}$** |
| 1IRA | - / 1 | 17.5 | 38.6 | 0.04 | 0.94 |
| 1H1V | - / 1 | 11.9 | 31.1 | 0.08 | 0.94 |
| 1Y64 | - / 1 | 10.3 | 32.2 | 0.07 | 0.92 |
| 1F6M | - / 1 | 14.1 | 39.4 | 0.00 | 1.00 |
| 1FAK | - / 1 | 11.4 | 28.5 | 0.01 | 0.99 |
| 1ZLI | - / 1 | 14.8 | 25.3 | 0.02 | 0.99 |
| 1E4K | ★ / 1 | 4.1 | 9.8 | 0.58 | 0.58 |
| 1IBR | - / 1 | 9.6 | 30.9 | 0.11 | 0.90 |
| 1KKL | ★ / 1 | 3.1 | 12.3 | 0.56 | 0.60 |
| 1NPE | ★★ / 1 | 1.7 | 6.1 | 0.85 | 0.45 |
| 1DFJ | ★★ / 116 | 1.8 | 5.6 | 0.63 | 0.59 |

[a]The quality is expressed according to CAPRI criteria (see **Experimental Procedures**). The reported rank is the rank of the individual models prior to clustering. See also **Table S1-S3**, [b]The ranking is based on the HADDOCK Score: 1.0 $E_{vdw}$ + 0.2 $E_{elec}$ + 1.0 $E_{desol}$(see **Main text**), [c]When no acceptable solution is generated, the values for the top ranked structure are reported.

Besides providing acceptable-to-good solutions for all of the cases, FMD-HADDOCK could rank them at the top using the standard scoring scheme in HADDOCK consisting of a weighted sum of van der Waals and electrostatic energies and an empirical desolvation term [76](HADDOCK$_{score}$ = 1.0 $E_{vdw}$ + 0.2

$E_{elec}$ + 1.0 $E_{desol}$). Furthermore, the quality of the models improved for 1E4K, 1KKL, 1NPE, and 1DFJ, for which two-body docking already provided reasonable solutions. 1E4K's and 1KKL's best models, ranked among the top five, are now a two star (good) solution and 1NPE's top ranking structure is almost a three star (high accuracy) prediction (with i-RMSD=1.1Å and Fnat=0.95). For 1DFJ the quality of the best model did best predictions superimposed onto the reference crystal structures are shown in **Figure 4**.

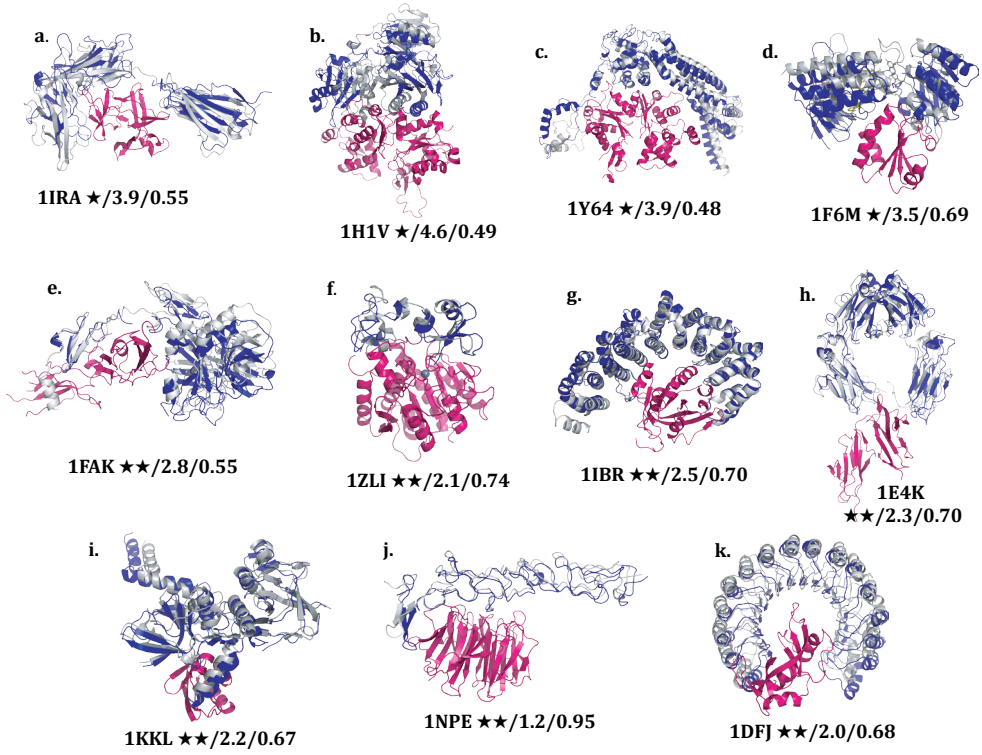



**Figure 4: View of the best HADDOCK solutions superimposed onto the respective reference crystal structures.** The docked complexes are shown in magenta (ligand) and blue (receptor) and the reference crystal structure (only the receptor is shown) in grey: **a.** 1IRA **b.** 1H1V **c.** 1Y64 **d.** 1F6M **e.** 1FAK **f.** 1ZLI **g.** 1IBR **h.** 1E4K **i.** 1KKL **j.** 1NPE **k.** 1DFJ. The quality of the models is indicated as: number of stars / interface RMSD (in Å) / fraction of native contacts (See **Experimental Procedures**). The figures were generated with Pymol [74].

**Table 4. Cluster statistics of *Flexible Multidomain Docking*.**[a]

| PDB ID | Quality / Rank | i-RMSD (Å) | l-RMSD (Å) | $F_{nat}$ | $F_{nonnat}$ |
|---|---|---|---|---|---|
| 1IRA | ★ / 1 | 4.2 ± 0.4 | 8.0 ± 1.0 | 0.54 ± 0.03 | 0.64 ± 0.07 |
| 1H1V | ★ / 2 | 4.8 ± 0.2 | 10.4 ± 1.2 | 0.50 ± 0.06 | 0.66 ± 0.06 |
| 1Y64 | ★ / 1 | 4.8 ± 0.6 | 11.3 ± 1.0 | 0.44 ± 0.06 | 0.74 ± 0.03 |
| 1F6M | ★ / 1 | 3.1 ± 0.3 | 9.2 ± 0.8 | 0.49 ± 0.08 | 0.64 ± 0.04 |
| 1FAK | ★ / 2 | 3.7 ± 0.5 | 9.8 ± 4.5 | 0.48 ± 0.02 | 0.61 ± 0.05 |
| 1ZLI | ★★ / 1 | 2.2 ± 0.2 | 3.8 ± 0.4 | 0.77 ± 0.03 | 0.47 ± 0.02 |
| 1E4K | ★ / 1 | 2.5 ± 0.2 | 5.9 ± 0.1 | 0.64 ± 0.05 | 0.46 ± 0.04 |
| 1IBR | ★★ / 1 | 2.7 ± 0.6 | 5.1 ± 1.1 | 0.67 ± 0.04 | 0.57 ± 0.03 |
| 1KKL | ★★ / 1 | 2.4 ± 0.2 | 5.4 ± 0.5 | 0.69 ± 0.05 | 0.56 ± 0.03 |
| 1NPE | ★★ / 2 | 2.2 ± 1.1 | 6.5 ± 4.7 | 0.82 ± 0.22 | 0.46 ± 0.06 |
| 1DFJ | ★ / 1 | 3.2 ± 0.8 | 9.1 ± 0.8 | 0.59 ± 0.02 | 0.67 ± 0.02 |

[a]The quality is expressed according to the CAPRI criteria (see **Experimental Procedures**). The rank corresponds to the cluster ranking based on the average HADDOCK score (1.0 $E_{vdw}$ + 0.2 $E_{elec}$ + 1.0 $E_{desol}$) of the top 4 members of a cluster. RMSD's and Fnat are reported as averages ± standard deviation calculated over the top 4 members of a cluster.

**Table 5. Number of docking models of various qualities for the various stages of the *Flexible Multidomain Docking* protocol. [a]**

| | Quality (★ / ★★ / ★★★) | | |
|---|---|---|---|
| **PDB ID** | ***it0***[a] | ***it1***[a] | ***water***[a] |
| 1IRA | 27 / - / - | 43 / - / - | 44 / - / - |
| 1H1V | 50 / - / - | 108 / - / - | 116 / - / - |
| 1Y64 | - / - / - | 1 / - / - | 1 / - / - |
| 1F6M | 2096 / - / - | 382 / - / - | 382 / - / - |
| 1FAK | 151 / 3 / - | 70 / 11 / - | 71 / 12 / - |
| 1ZLI | 1496 / 3246 / - | 5 / 395 / - | 3 / 397 / - |
| 1E4K | 2569 / 1371 / - | 87 / 212 / - | 98 / 201 / - |
| 1IBR | 1287 / - / - | 195 / 104 / - | 196 / 103 / - |
| 1KKL | 521 / 1844 / - | 323 / 77 / - | 304 / 96 / - |
| 1NPE | 984 / 32 / 1 | 356 / 26 / - | 353 / 30 / - |
| 1DFJ | 210 / - / - | 96 / - / - | 116 / 1 / - |

[a] *it0:* Rigid Multidomain docking, *it1:* Semi-flexible refinement in torsion angle space, *water:* Final explicit solvent refinement.

Within the standard HADDOCK protocol the final ranking of solutions is based on cluster averages rather than individual ranks. Each of the top ranking clusters contained at least one acceptable prediction, with an average Fnat of 0.44 for the worst and of 0.82 for the best case (**Table 4**). This stresses the excellent performance of our scoring scheme. We also analyzed the number of acceptable or better structures generated during each stage of the docking (**Table 5**). One can observe a clear decrease in the number of acceptable solutions as a function of the docking difficulty (related to the extent of the conformational changes). For the challenging cases, scoring becomes even more critical since the fraction of good solutions within the pool of sampled conformations decreases, as can be seen in the case of 1Y64: For this complex, only one acceptable solution is generated after refinement; despite this, HADDOCK ranked it at the top (**Table 3**). For the other cases, the FMD protocol provided a pool of near native predictions after each refinement step (*it1, water*) (**Table 5**).



**4. Discussion**

Docking has become a popular approach to model biomolecular interactions. Current approaches are able to deal with small side-chain and backbone alterations while dealing with large conformational changes is still one of the major bottlenecks in the field. Considering the various types of conformational changes and the focus of existing docking approaches on mainly dealing with rather local small changes, it is evident that new developments are required to tackle modeling of large conformational changes occurring upon binding.

Here we have presented a straightforward and easy-to-apply *Flexible Multidomain Docking* protocol that can deal with large collective backbone conformational changes by treating the flexible partner as a collection of sub-domains with connectivity restraints between them. Our results have revealed that FMD outperforms standard two-body docking especially in the cases with conformational change range ≥3Å (1IRA, 1H1V, 1Y64, 1F6M, 1FAK, 1ZLI), something that has never been demonstrated before. Even in the case of smaller conformational changes (<3Å) the FMD protocol improves in general both the quality and ranking of the solutions upon two-body docking (1E4K, 1IBR, 1KKL, 1NPE, 1DFJ). Interestingly two-body docking worked well for 1E4K but not for 1IBR, although both receptors experience the same extent of conformational change (2.9Å). Possible explanations for this could be that: (i) The domain motion of 1E4K can be explained with only one hinge while more than one are needed for 1IBR and (ii) the interface area of 1IBR is twice as large as that of 1E4K (1685Å$^2$ against 810Å$^2$). The performance of standard docking methods seems to be affected not only by the

amount of conformational changes but also by their complexity and the extent of the interface to be predicted.

Other factors that may affect the docking performance are the size of the sub-domains to be docked and the number of hinges at which a receptor is cut. Our results revealed that FMD is robust with respect to the size of the segments involved in docking. For 1Y64, 1ZLI, 1KKL and 1NPE one of the sub-domains consists of less than fifty residues while the others have larger domains. Shorter segments are more prone to be flexible, thus making it harder to model the correct binding mode. Also our protocol is performing well when more than one hinge is involved in the domain motion as illustrated by the 1F6M case. FMD is quite successful to model a wide range of conformational changes (from 1.5Å to 20Å), in cases when only two domains are involved. It does not perform as well when the system complexity increases like for example for the former CAPRI target, 1TLV [77], which consists of a homodimer with each monomer experiencing a large conformational change of 12.7Å upon activation. In this case, a four body docking failed to generate any acceptable solution (best i-RMSD 6.2Å and global RMSD 10.3 Å), where the multi-stage docking protocol of MolFit [28] could generate a model with a global RMSD of 5.6Å.



Another parameter that affects the docking quality is the interface information used to drive docking. In order to focus on our ability to deal with conformational changes, we assumed in this work perfect interface information; this is a best-case scenario and one can therefore expect that the performance will depend on the quality and reliability of the supplied data. We investigated this aspect by using bioinformatics interface predictions obtained from our in-house consensus predictor, CPORT, to drive docking (http://haddock.chem.uu.nl/services/CPORT/) [39]. We could only generate acceptable solutions for 1FAK, 1ZLI and 1KKL (**Table S1**). This is directly related to the specificity of the prediction (see **Table S2**). Only for four cases (1FAK, 1ZLI, 1KKL and 1DFJ) the specificity of CPORT predictions was above 30% for both receptor and ligand and for three of these acceptable to medium quality solutions could be obtained. This is in line with what was observed in systematic HADDOCK runs, where CPORT was used to drive docking [39].

In general, our FMD protocol could generate at least an acceptable solution for each case, and even medium-quality predictions for seven of them. The fraction of native contacts for the best models was above 0.5 (the CAPRI threshold for high quality predictions) for nine of them. The strength of FMD-HADDOCK resides mainly in its ability to deal with large-scale and small, induced conformational changes at the same time.

Our benchmark shares common cases with three other hybrid methodologies (see **Introduction**): 1NPE and 1IBR with FlexDock (see **Table 2** in

Schneidman-Duhovny *et al.* [29]); 1DFJ and 1IBR with ATTRACT (see **Table 5** in May *et al.* [25]) and FiberDock (**Table 4** in Mashiach *et al.* [32]). Comparing the best l-RMSDs for 1IBR, FlexDock, FiberDock and FMD produce similar somewhat better results than ATTRACT. FMD however outperforms FlexDock in scoring. A similar situation is observed for 1NPE. In the case of 1DFJ FMD, FiberDock and ATTRACT perform well in both sampling and scoring, although FiberDock's best model has a better l-RMSD.

The excellent performance of the new FMD-HADDOCK protocol represents already a major step in dealing with large conformational changes occurring upon binding. One question however remains: How can we predict which type of conformational change should be expected upon binding so that one can choose the best suited method? Dobbins *et al.* [24] pointed out that Normal Mode Analysis can provide some discrimination among the various types of conformational changes. They stated that the slowest mode of a protein experiencing significant conformational changes (Cα-RMSD>2.0Å) has a 2.5 times lower frequency than for proteins undergoing limited conformational change (Cα-RMSD<1.0Å). To further investigate this issue, we considered a set of 268 proteins experiencing different levels of conformational changes (0.3Å≤Cα-RMSD≤19.5Å). These were taken from the Docking Benchmark 4.0 [73], excluding multimers, antibody-antigen complexes and proteins having less than fifty amino acids. For all single proteins, eigenvalues from a Gaussian Network Model (GNM, see **Experimental Procedures**) [37] were obtained from the HingeProt web server [36]. They were normalized and their cumulative sum (ranging from 0 to 1) was calculated and plotted as a function of the number of modes. This plot can be interpreted as *the fraction of motion explained* by a given number of modes. Here we assumed that proteins undergoing larger conformational changes have smaller GNM-eigenvalues, corresponding to collective low frequency motions. For these, the cumulative sum of the normalized GNM-eigenvalues should increase slower as a function of the number of modes than for rigid cases, in agreement with the observations of Dobbins *et al.* [24] (**Figure S1**). To test this assumption, we developed a simple predictor, which predicts proteins as being *rigid* or *flexible* based on the fraction of motion explained by a given number of modes (see **Supplementary Experimental Procedures**). This classifier performs with the best accuracy (65±16%, cross-validated accuracy) on the set of 268 proteins when a 2.6Å RMSD cut-off is used to classify these proteins as *rigid* and *flexible* and the cut-off for the *fraction of motion explained* for the first 50 modes is taken as 0.08 (i.e. the cumulative sum of the eigenvalues for the first 50 modes should be < 0.08 for the protein to be predicted as *flexible*). Our simple predictor was able to classify the receptors of the FMD benchmark with 64% accuracy (with seven out of nine

*flexible* cases and one out of two *rigid* cases being correctly classified, see **Figure 5**). The cumulative sum of the normalized eigenvalues for a given number of modes may thus be deterministic in predicting the range of the conformational change that could be expected. This could provide a mean to select an appropriate method to deal with different types of expected conformational changes.

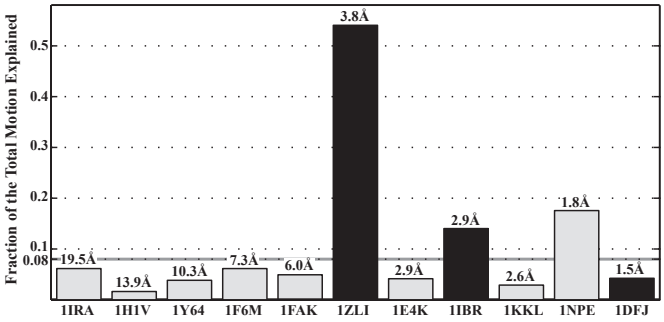



**Figure 5: Fraction of motion explained by the first 50 GNM modes for the 11 receptors from FMD Benchmark.** Grey bars indicate correctly classified; black bars indicate the misclassified cases. The extent of the conformational change is indicated on top of each bar.

## 5. Conclusions

We have developed a new *Flexible Multidomain Docking* protocol that follows a "divide-and-conquer" approach to model large-scale domain motions and small- to medium-scale rearrangements within an interface at the same time. This represents a major step in dealing with one of the major challenges and limitations in the modeling of biomolecular complexes. Our analysis also identified indicators that might allow us to predict the extent of the conformational changes and thus select the most appropriate method to deal with them. This remains, however, a challenging problem as a variety of approaches will still have to be combined in order to properly describe simultaneously all possible types of changes. The availability of some experimental information will become crucial to drive the docking and/or for validation of the resulting models as the complexity of the problem increases. As a final remark we note that *Flexible Multidomain Docking* as implemented in HADDOCK 2.1 is now also accessible via the multibody docking interface of the HADDOCK web server [34].

## Supplementary Information

Here are the details of this section's content and how is relates to the rest of the chapter:

**Figure S1** (Cumulative Sum of the Normalized GNM-Eigenvalues as a function of the first 50 modes) is related to **Figure 5**. The data values provided in **Figure 5** correspond to the *Cumulative Sum of the Normalized GNM-Eigenvalues* of the red lines at *Number of modes*=50. **Table S1** is related to **Table 3**. In **Table S1**, we present the performance of our protocol when bioinformatics predictions are used to drive the docking. In the main text these results are discussed and compared with the results given in Table 3. **Table S2** is related to **Table 3**. **Table S2** is also directly related to **Table S1** since it provides the quality assessment of the bioinformatics predictions used in docking. **Table S3** is related to **Table 3**. The data given in **Table S3** are used in the FMD-HADDOCK procedure to drive the docking, the results of which are presented in **Table 3**. **Supplementary Experimental Procedures** is related to **Figure 5**, which illustrates the performance of our classifier in predicting the extent of the conformational changes. The **Supplemental Experimental Procedures** section provides a description of how our classifier was developed and validated.

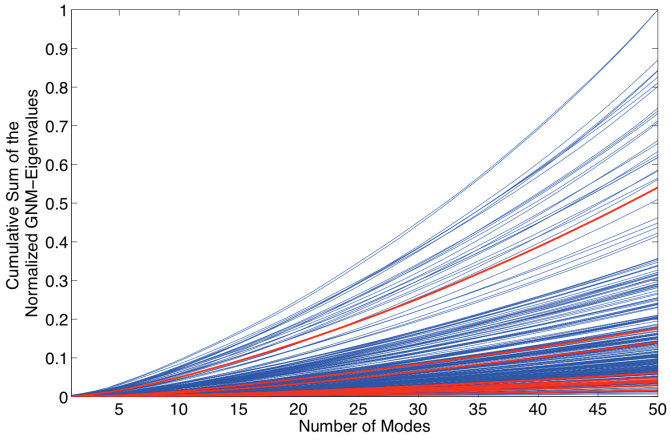



**Figure S1. Cumulative Sum of the Normalized GNM-Eigenvalues as a function of the first 50 modes.** The red lines correspond to the receptors of the FMD benchmark, while the blue lines indicate the remaining eligible cases from the Docking Benchmark 4.0 [73]. Eight out of eleven receptors (1IRA, 1H1V, 1F6M, 1Y64, 1FAK, 1E4K, 1KKL, 1DFJ) have particularly small slopes compared to the other the cases.

**Table S1. CPORT-driven FMD-Benchmark docking results.**

| PDB ID | Quality / Rank | i-RMSD (Å) | Best i-RMSD (Å) / Rank | l-RMSD (Å) | $F_{nat}$ | $F_{nonnat}$ |
|---|---|---|---|---|---|---|
| 1IRA | - / 1 | 13.8 | 8.1 / 102 | 32.7 | 0.00 | 1.00 |
| 1H1V | - / 1 | 21.6 | 16.4 / 318 | 57.2 | 0.00 | 1.00 |
| 1Y64 | - / 1 | 21.4 | 9.4 / 205 | 52.6 | 0.00 | 1.00 |
| 1F6M | - / 1 | 12.8 | 10.4 / 155 | 25.3 | 0.03 | 0.97 |
| 1FAK | ★ / 61 | 3.4 | 3.4 / 61 | 10.8 | 0.23 | 0.55 |
| 1ZLI | ★ / 5 | 4.2 | 4.2 / 5 | 9.2 | 0.34 | 0.5 |
| 1E4K | - / 1 | 24.2 | 11.2 / 76 | 66.8 | 0.00 | 1.00 |
| 1IBR | - / 1 | 19.7 | 11.1 / 70 | 51.9 | 0.00 | 1.00 |
| 1KKL | ★★ / 154 | 1.9 | 1.9 / 154 | 5.3 | 0.70 | 0.26 |
| 1NPE | - / 1 | 11.4 | 4.5 / 138 | 35.1 | 0.00 | 1.00 |
| 1DFJ | - / 1 | 16.9 | 5.1 / 217 | 59.4 | 0.00 | 1.00 |

**Table S2. Quality assessment of the CPORT predictions.** (Cases for which CPORT-driven FMD-HADDOCK could generate at least one acceptable solution are indicated in bold).

| PDB ID | Specificity (%) | Sensitivity (%) |
|---|---|---|
| 1IRA | 21.1 / 43.2 | 31.9 / 44.2 |
| 1H1V | 8.7 / 17.7 | 14.8 / 51.8 |
| 1Y64 | 8.4 / 20.7 | 27.1 / 37.2 |
| 1F6M | 9.7 / 43.5 | 29.2 / 80.0 |
| **1FAK** | **42.8 / 40.9** | **57.1 / 37.5** |
| **1ZLI** | **53.1 / 55.3** | **60.7 / 63.6** |
| 1E4K | 3.6 / 19.6 | 9.1 / 52.6 |
| 1IBR | 22.4 / 26.1 | 31.7 / 36.2 |
| **1KKL** | **30.0 / 31.6** | **53.6 / 71.4** |
| 1NPE | 15.0 / 10.7 | 25.0 / 22.2 |
| 1DFJ | 48.2 / 47.5 | 69.2 / 57.6 |

***Supplementary Experimental Procedures:***

*Predicting the amount of conformational change*

In order to see if GNM-eigenvalues have any value in predicting the amount of conformational changes and identify large domain motions we analyzed a set of 268 proteins experiencing different levels of conformational changes (0.3Å≤Cα-RMSD≤19.5Å). These were taken from the Docking Benchmark 4.0 [73] excluding multimers, antibody-antigen complexes and proteins having less than fifty residues.

For all single proteins, GNM-Eigenvalues [37] were downloaded from the HingeProt webserver [36]. They were normalized and their cumulative sum (ranging from 0 to 1) was calculated and plotted as a function of the number of modes. This plot will be referred to as *the fraction of motion explained* by a given number of modes.

We designed a simple predictor for classifying the proteins as *rigid* or *flexible* based on the value of the cumulative sum for a given number of modes. The parameters of our predictor were optimized in a grid search to maximize the prediction accuracy on the set of 268 proteins described above. The following parameters were considered: (i) *RMSDcut-off* for classifying the proteins as *rigid* and *flexible*, (ii) number of modes considered, *Nmodes* (from 1 to 50 is steps of 5) and (iii) cutoff for the fraction of motion explained, *FMcut-off* (from 0.01 to 0.20 in steps of 0.01). We limited the RMSD cutoff between 2 and 3Å, 2Å being the limit defining the challenging cases in the protein Benchmark 4.0.



The optimization was performed using cross-validation by dividing randomly the data into ten sub-sets and using nine of them for the optimization. The accuracy was measured both on the training and left out sets. This procedure was carried out ten times so that all of the sub-sets were used once for validation purposes.

For a given set of parameters (*RMSDcut-off*, *Nmodes*, *FMcut-off*) the decision process of our predictor is very simple:

- Cumulative sum for *Nmodes* > *FMcutoff value* ➔ rigid protein
- Cumulative sum for *Nmodes* <= *FMcutoff value* ➔ flexible protein

The prediction's accuracy was defined as:

$$\text{Accuracy} = \tfrac{1}{2}\,(\text{Accuracy}_{\text{rigid}} + \text{Accuracy}_{\text{flexible}})$$

where the rigid and flexible classes are defined based on the chosen *RMSDcut-off*. For each class the accuracy is simply the number of correct predictions divided by the total number of members in that class. We chose for this class-specific accuracy evaluation to remove size effects since the rigid class typically contains many more members than the flexible class. With this measure a random predictor would give an overall accuracy of 50%.

The cross-validated average accuracy obtained for ten different test sets is 65±16%; it was obtained for the following combination of parameters: *FMcut-off* = 0.08, *Nmodes* = 50 and *RMSDcut-off* = 2.6Å, which were consistently found over the ten different optimization set.

**Table S3. AIRs used in the docking of the benchmark cases.** The residue numbering corresponds to the numbering of the unbound receptor and ligand.

| AIR used for docking | | |
|---|---|---|
| **ID** | **Receptor** | **Ligand** |
| 1IRA | 6,8-13,25,27-30,105-113,117-121,123, 124,126,160,198,200,202,233- | 8-11,14,16,18-20,22,24-27,29,31,33-40,42, 43,50-54,56,102,107,126-130,147,149-151 |
| 1H1V | 428,429,465,467,473-478,480,482,484, | 543-548,567-569,621,696,711,714,715, 718,726,734,741,742,745,746,748-751, |
| 1Y64 | 1359,1360,1362,14021409,1411,1412, 1414,1415,1424,1427,1428,1430- | 4,5,6,99,100,102,124,125,128,130,143-148,167,225,318,319,321- |
| 1F6M | 37,39,44,81,83,85,99,100,128,131,133, 137-139,141-143,215-217,237,238 | 31-35,37,40,41,44,60,67,70-75,77,91,93-98 |
| 1FAK[a] | 1,2,4,5,8,9,27-30,34-39,41-47,50,53,55, | 17,18,20,22,24,37,39,40-48,50,51,56,58,61, 74,76,90-96,109,110,112,130-133,135,140, |
| 1ZLI | 1,4,5,9,10,11,12,15,26-29,34,37,41-46, 52-55,57,72-74 | 69,71-73,119-125,127,145,155-157,161-164, |
| 1E4K[a] | 9,10,11,37,38,39,41,69,70,71,99,223-227,314-318,320 | 88-90,113,116,117,119,120,129,130-132, 135,155,158-161 |
| 1IBR | 12,45-47,64,70,74-79,81,82,103,104, 106,107,109,110-114,137,139,140- | 10,11,12,13,15,18,22,25,51,52,55, 56,59,60,62,63,67-69,72,105-107, |
| 1KKL[a] | 2,3,4,6,23,44-46,48,64,65,70,76,312, 313,316,317,319,320 | 12-16,40-49,51-57 |
| 1NPE | 27,29-33,37-40,42-47,73-78,82,89 | 9,32,34-37,39,55,56,80,82,98,99,125,141, 143,168,170,186,210,212,226,248,249,251, |
| 1DFJ | 6,7,31,32,60,89,117,146,202,228,257, 259,283,285,314,316,342,379,397,399, | 4,7,11,12,23,24,28,31,32,35,38-44,65-67,69, |

[a]For these cases, the receptor, consisting of multiple chains, was renumbered starting from one and continuously over the various chains (for docking a single chainID was used).

## References

[1] A.M.J.J. Bonvin, Flexible protein-protein docking, Curr. Opin. Struct. Biol. 16 (2006) 194–200.
[2] M.F. Lensink, R. Mendez, Recognition-induced conformational changes in protein-protein docking, Curr. Pharm. Biotechnol. 9 (2008) 77–86.
[3] I.S. Moreira, P.A. Fernandes, M.J. Ramos, Protein-protein docking dealing with the unknown., J. Comput. Chem. 31 (2010) 317–42.
[4] C. Pons, S. Grosdidier, A. Solernou, L. Pérez-Cano, J. Fernández-Recio, Present and future challenges and limitations in protein-protein docking., Proteins: Struct., Funct., Bioinf. 78 (2010) 95–108.
[5] D.W. Ritchie, Recent progress and future directions in protein-protein docking, Curr. Protein Pept. Sci. 9 (2008) 1–15.
[6] M. Zacharias, Accounting for conformational changes during protein-protein docking., Curr. Opin. Struct. Biol. 20 (2010) 180–6.
[7] N. Andrusier, E. Mashiach, R. Nussinov, H.J. Wolfson, Principles of flexible protein-protein docking, Proteins: Struct., Funct., Bioinf. 73 (2008) 271–289.
[8] E. Fischer, No Title, in: Einfluss Der Configuration Auf Die Wirkung Der Enzyme, 1894: pp. 2985–2993.
[9] D.E. Koshland, Application of a Theory of Enzyme Specificity to Protein Synthesis, Proc. Natl. Acad. Sci. U. S. A. 44 (1958) 98–104.
[10] S. Kumar, B. Ma, C.J. Tsai, N. Sinha, R. Nussinov, Folding and binding cascades: dynamic landscapes and population shifts, Protein Sci. 9 (2000) 10–19.
[11] R. Grunberg, M. Nilges, J. Leckner, Flexibility and conformational entropy in protein-protein binding, Structure. 14 (2006) 683–693.
[12] M. Gerstein, W. Krebs, A database of macromolecular motions, Nucleic Acids Res. 26 (1998) 4280–4290.
[13] M. Gerstein, N. Echols, Exploring the range of protein flexibility, from a structural proteomics perspective, Curr. Opin. Chem. Biol. 8 (2004) 14–19.
[14] T. Mittag, L.E. Kay, J.D. Forman-Kay, Protein dynamics and conformational disorder in molecular recognition, J. Mol. Recognit. 23 (2010) 105–116.
[15] O. Bachar, D. Fischer, R. Nussinov, H. Wolfson, A computer vision based technique for 3-D sequence-independent structural comparison of proteins, Protein Eng. 6 (1993) 279–288.
[16] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, I.A. Vakser, Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 2195–2199.
[17] C. Dominguez, R. Boelens, A.M.J.J. Bonvin, HADDOCK: a protein-protein docking approach based on biochemical or biophysical information., J. Am. Chem. Soc. 125 (2003) 1731–7.
[18] J.J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, et al., Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations., J. Mol. Biol. 331 (2003) 281–99.
[19] S. Chaudhury, J.J. Gray, Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles, J. Mol. Biol. 381 (2008) 1068–1087.
[20] M. Król, R.A.G. Chaleil, A.L. Tournier, P.A. Bates, M. Krol, Implicit flexibility in protein docking: cross-docking and local refinement, Proteins: Struct., Funct., Bioinf. 69 (2007) 750–7.
[21] K. Bastard, A. Thureau, R. Lavery, C. Prevost, Docking macromolecules with flexible segments, J. Comput. Chem. 24 (2003) 1910–1920.
[22] A. May, M. Zacharias, Protein-protein docking in CAPRI using ATTRACT to account for global and local flexibility, Proteins: Struct., Funct., Bioinf. 69 (2007) 774–780.

[23] R. Das, I. André, Y. Shen, Y. Wu, A. Lemak, S. Bansal, et al., Simultaneous prediction of protein folding and docking at high resolution., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 18978–83.
[24] S.E. Dobbins, V.I. Lesk, M.J.E. Sternberg, Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 10390–10395.
[25] A. May, M. Zacharias, Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking., Proteins: Struct., Funct., Bioinf. 70 (2008) 794–809.
[26] X. Li, I.H. Moal, P.A. Bates, Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding, Proteins: Struct., Funct., Bioinf. 78 (2010) 3189–3196.
[27] I.H. Moal, P.A. Bates, SwarmDock and the Use of Normal Modes in Protein-Protein Docking, I. J. Mol. Sci. 11 (2010) 3623–3648.
[28] E. Ben-Zeev, N. Kowalsman, A. Ben-Shimon, D. Segal, T. Atarot, O. Noivirt, et al., Docking to single-domain and multiple-domain proteins: old and new challenges, Proteins: Struct., Funct., Bioinf. 60 (2005) 195–201.
[29] D. Schneidman-Duhovny, R. Nussinov, H.J. Wolfson, Automatic prediction of protein interactions with large scale motion, Proteins: Struct., Funct., Bioinf. 69 (2007) 764–773.
[30] C. Wang, P. Bradley, D. Baker, Protein-protein docking with backbone flexibility, J. Mol. Biol. 373 (2007) 503–519.
[31] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, H.J. Wolfson, PatchDock and SymmDock: servers for rigid and symmetric docking, Nucleic Acids Res. 33 (2005) W363–7.
[32] E. Mashiach, R. Nussinov, H.J. Wolfson, FiberDock: Flexible induced-fit backbone refinement in molecular docking, Proteins: Struct., Funct., Bioinf. 78 (2010) 1503–1519.
[33] S.J. de Vries, A.D.J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, et al., HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets, Proteins: Struct., Funct., Bioinf. 69 (2007) 726–733.
[34] S.J. de Vries, M. van Dijk, A.M.J.J. Bonvin, The HADDOCK web server for data-driven biomolecular docking, Nature Protoc. 5 (2010) 883–897.
[35] E. Karaca, A.S. Melquiond, S.J. de Vries, P.L. Kastritis, A.M. Bonvin, Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server, Mol. Cell. Biol. 9 (2010) 1784–1794.
[36] U. Emekli, D. Schneidman-Duhovny, H.J. Wolfson, R. Nussinov, T. Haliloglu, HingeProt: automated prediction of hinges in protein structures, Proteins: Struct., Funct., Bioinf. 70 (2008) 1219–1227.
[37] T. Haliloglu, I. Bahar, B. Erman, Gaussian dynamics of folded proteins, Phys. Rev. Lett. 79 (1997) 3090–3093.
[38] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model, Biophys. J. 80 (2001) 505–515.
[39] S.J. de Vries, A.M.J.J. Bonvin, CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK, PLoS ONE. 6 (2011) e17695.
[40] R. Mendez, R. Leplae, L. De Maria, S.J. Wodak, R. Méndez, Assessment of blind predictions of protein-protein interactions: current status of docking methods, Proteins: Struct., Funct., Bioinf. 52 (2003) 51–67.
[41] H. Schreuder, C. Tardif, S. Trump-Kallmeyer, A. Soffientini, E. Sarubbi, A. Akeson, et al., A new cytokine-receptor binding mode revealed by the crystal structure of the IL-1 receptor with an antagonist., Nature. 386 (1997) 194–200.

[42] G.P. Vigers, D.J. Dripps, C.K. Edwards 3rd, B.J. Brandhuber, X-ray crystal structure of a small antagonist peptide bound to interleukin-1 receptor type 1, J. Biol. Chem. 275 (2000) 36927–36933.
[43] H.A. Schreuder, J.M. Rondeau, C. Tardif, A. Soffientini, E. Sarubbi, A. Akeson, et al., Refined crystal structure of the interleukin-1 receptor antagonist. Presence of a disulfide link and a cis-proline, Eur. J. Biochem. 227 (1995) 838–847.
[44] H. Choe, L.D. Burtnick, M. Mejillano, H.L. Yin, R.C. Robinson, S. Choe, The calcium activation of gelsolin: insights from the 3A structure of the G4-G6/actin complex, J. Mol. Biol. 324 (2002) 691–702.
[45] L.D. Burtnick, E.K. Koepf, J. Grimes, E.Y. Jones, D.I. Stuart, P.J. McLaughlin, et al., The crystal structure of plasma gelsolin: implications for actin severing, capping, and nucleation, Cell. 90 (1997) 661–670.
[46] M.R. Bubb, L. Govindasamy, E.G. Yarmola, S.M. Vorobiev, S.C. Almo, T. Somasundaram, et al., Polylysine induces an antiparallel actin dimer that nucleates filament assembly: crystal structure at 3.5-A resolution, J. Biol. Chem. 277 (2002) 20999–21006.
[47] T. Otomo, D.R. Tomchick, C. Otomo, S.C. Panchal, M. Machius, M.K. Rosen, Structural basis of actin filament nucleation and processive capping by a formin homology 2 domain, Nature. 433 (2005) 488–494.
[48] Y. Xu, J.B. Moseley, I. Sagot, F. Poy, D. Pellman, B.L. Goode, et al., Crystal structures of a Formin Homology-2 domain reveal a tethered dimer architecture, Cell. 116 (2004) 711–723.
[49] S.A. Rizvi, V. Tereshko, A.A. Kossiakoff, S.A. Kozmin, Structure of bistramide A-actin complex at a 1.35 angstroms resolution, J. Am. Chem. Soc. 128 (2006) 3882–3883.
[50] B.W. Lennon, C.H. Williams Jr, M.L. Ludwig, Twists in catalysis: alternating conformations of Escherichia coli thioredoxin reductase, Science. 289 (2000) 1190–1194.
[51] B.W. Lennon, C.H. Williams Jr, M.L. Ludwig, Crystal structure of reduced thioredoxin reductase from Escherichia coli: structural flexibility in the isoalloxazine ring of the flavin adenine dinucleotide cofactor, Protein Sci. 8 (1999) 2366–2379.
[52] M. Nikkola, F.K. Gleason, J.A. Fuchs, H. Eklund, Crystal structure analysis of a mutant Escherichia coli thioredoxin in which lysine 36 is replaced by glutamic acid, Biochemistry. 32 (1993) 5093–5098.
[53] E. Zhang, R. St Charles, A. Tulinsky, Structure of extracellular tissue factor complexed with factor VIIa inhibited with a BPTI mutant, J. Mol. Biol. 285 (1999) 2089–2104.
[54] A.C. Pike, A.M. Brzozowski, S.M. Roberts, O.H. Olsen, E. Persson, Structure of human factor VIIa and its implications for the triggering of blood coagulation, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 8925–8930.
[55] M. Huang, R. Syed, E.A. Stura, M.J. Stone, R.S. Stefanko, W. Ruf, et al., The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor, Fab 5G9 and TF.G9 complex, J. Mol. Biol. 275 (1998) 873–894.
[56] J.L. Arolas, G.M. Popowicz, J. Lorenzo, C.P. Sommerhoff, et al., The three-dimensional structures of tick carboxypeptidase inhibitor in complex with A/B carboxypeptidases reveal a novel double-headed binding mode, J. Mol. Biol. 350 (2005) 489–498.
[57] D. Pantoja-Uceda, J.L. Arolas, P. Garcia, E. Lopez-Hernandez, D. Padro, F.X. Aviles, et al., The NMR structure and dynamics of the two-domain tick carboxypeptidase inhibitor reveal flexibility in its free form and stiffness upon binding to human carboxypeptidase B, Biochemistry. 47 (2008) 7066–7078.
[58] P.J. Barbosa Pereira, S. Segura-Martin, B. Oliva, C. Ferrer-Orta, F.X. Aviles, M. Coll, et al., Human procarboxypeptidase B: three-dimensional structure and implications for thrombin-activatable fibrinolysis inhibitor (TAFI), J. Mol. Biol. 321 (2002) 537–547.
[59] P. Sondermann, R. Huber, V. Oosthuizen, U. Jacob, The 3.2-A crystal structure of the human IgG1 Fc fragment-Fc gammaRIII complex, Nature. 406 (2000) 267–273.

[60] S. Matsumiya, Y. Yamaguchi, J. Saito, M. Nagano, H. Sasakawa, S. Otaki, et al., Structural comparison of fucosylated and nonfucosylated Fc fragments of human immunoglobulin G1, J. Mol. Biol. 368 (2007) 767–779.
[61] Y. Zhang, C.C. Boesen, S. Radaev, A.G. Brooks, W.H. Fridman, C. Sautes-Fridman, et al., Crystal structure of the extracellular domain of a human Fc gamma RIII, Immunity. 13 (2000) 387–395.
[62] I.R. Vetter, A. Arndt, U. Kutay, D. Gorlich, A. Wittinghofer, Structural view of the Ran-Importin beta interaction at 2.3 A resolution, Cell. 97 (1999) 635–646.
[63] R. Bayliss, T. Littlewood, M. Stewart, Structural basis for the interaction between FxFG nucleoporin repeats and importin-beta in nuclear trafficking, Cell. 102 (2000) 99–108.
[64] H.M. Kent, M.S. Moore, B.B. Quimby, A.M. Baker, A.J. McCoy, G.A. Murphy, et al., Engineered mutants in the switch II loop of Ran define the contribution made by key residues to the interaction with nuclear transport factor 2 (NTF2) and the role of this interaction in nuclear protein import, J. Mol. Biol. 289 (1999) 565–577.
[65] S. Fieulaine, S. Morera, S. Poncet, I. Mijakovic, A. Galinier, J. Janin, et al., X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 13437–13441.
[66] S. Fieulaine, S. Morera, S. Poncet, V. Monedero, V. Gueguen-Chaignon, A. Galinier, et al., X-ray structure of HPr kinase: a bacterial protein kinase with a P-loop nucleotide-binding domain, EMBO J. 20 (2001) 3917–3927.
[67] D.I. Liao, O. Herzberg, Refined structures of the active Ser83-->Cys and impaired Ser46-->Asp histidine-containing phosphocarrier proteins, Structure. 2 (1994) 1203–1216.
[68] J. Takagi, Y. Yang, J.H. Liu, J.H. Wang, T.A. Springer, Complex between nidogen and laminin fragments reveals a paradigmatic beta-propeller interface, Nature. 424 (2003) 969–974.
[69] J. Stetefeld, U. Mayer, R. Timpl, R. Huber, Crystal structure of three consecutive laminin-type epidermal growth factor-like (LE) modules of laminin gamma1 chain harboring the nidogen binding site, J. Mol. Biol. 257 (1996) 644–657.
[70] B. Kobe, J. Deisenhofer, A structural basis of the interactions between leucine-rich repeats and protein ligands, Nature. 374 (1995) 183–186.
[71] B. Kobe, J. Deisenhofer, Mechanism of ribonuclease inhibition by ribonuclease inhibitor protein based on the crystal structure of its complex with ribonuclease A, J. Mol. Biol. 264 (1996) 1028–1043.
[72] J. Nachman, M. Miller, G.L. Gilliland, R. Carty, M. Pincus, A. Wlodawer, Crystal structure of two covalent nucleoside derivatives of ribonuclease A, Biochemistry. 29 (1990) 928–937.
[73] H. Hwang, T. Vreven, J. Janin, Z. Weng, Protein-protein docking benchmark version 4.0, Proteins: Struct., Funct., Bioinf. 78 (2010) 3111–3114.
[74] W.L. DeLano, The PyMOL Molecular Graphics System on World Wide Web http://www.pymol.org, (2002).
[75] S. Hayward, Structural principles governing domain motions in proteins, Proteins: Struct., Funct., Bioinf. 36 (1999) 425–435.
[76] J. Fernández-Recio, M. Totrov, R. Abagyan, J. Fernandez-Recio, Identification of protein-protein interaction sites from docking energy landscapes, J. Mol. Biol. 335 (2004) 843–865.
[77] M. Graille, C.Z. Zhou, V. Receveur-Brechot, B. Collinet, N. Declerck, H. van Tilbeurgh, Activation of the LicT transcriptional antiterminator involves a domain swing/lock mechanism provoking massive structural changes, J. Biol. Chem. 280 (2005) 14780–14789.

Chapter

4

# On the Usefulness of Ion Mobility Mass Spectrometry and SAXS Data in Scoring Docking Decoys

Based on the research article:
**Ezgi Karaca** and **Alexandre M.J.J. Bonvin**, On the Usefulness of Ion Mobility Mass Spectrometry and SAXS Data in Scoring Docking Decoys, 2012, *submitted for publication*.


## Abstract

Scoring, the process of selecting the biologically relevant solution from a pool of generated conformations, is one of the major challenges in the field of biomolecular docking. A prominent way to cope with this challenge is to incorporate information-based terms into the scoring function. Within this context, we have integrated low-resolution shape data obtained from either Ion Mobility Mass Spectrometry (IMMS) or SAXS experiments, into the conventional scoring function of our information-driven docking program HADDOCK. Here, we systematically assess the strengths and weaknesses of IMMS- and SAXS-based scoring, either in isolation or in combination with the HADDOCK score. The results from an analysis of a large docking decoy set composed of dimers, generated by running HADDOCK in *ab inito* mode, reveal that the content of IMMS data is of too low resolution for selecting correct models, while, on the other hand, scoring with SAXS data leads to a significant performance improvement. The effectiveness of SAXS scoring however depends on the shape and arrangement of the complex with *prolate* and *oblate* systems showing the best performance. We observe that the highest accuracy is achieved when SAXS scoring is combined with the energy-based HADDOCK score.

## 1. Introduction

Understanding how a single cell functions is the fundamental quest of life sciences. This can only be comprehensively addressed once the structure-function relationships of biomolecular complexes occurring in that particular cell will have been explored thoroughly. There are two main classical experimental techniques that can reveal the structure of the biomolecular complexes in atomistic detail, X-Ray crystallography and NMR spectroscopy. Although these have helped immensely to shed light on the mechanical and functional world of biomolecules, they are faced with many challenges when the biomolecular systems under study become very large, comprise flexible or unstructured regions, exist in very tiny amounts, are membrane associated, or when their constituents interact only transiently [1,2]. While these limitations hamper acquiring high-resolution information low-resolution biochemical or biophysical data can often still be obtained. The disadvantage here is that, most of the time, these data are sparse and contain limited structural information compared to high-resolution methods. Therefore, computational modeling using integrative approaches, like macromolecular docking, are needed to translate these sparse low-resolution data into useful structural information [1,3].



Low-resolution spatial information can be derived from a variety of biophysical experimental techniques comprising NMR [3–5], cryo-EM [6,7], MS [8], IMMS [9,10], EPR [11], SAXS [1,4], FRET [12], and/or biochemical ones like mutagenesis and chemical footprinting [3,6,13]. The data obtained from these experiments can be integrated into a modeling procedure either during sampling i.e. *a priori* by restraining the conformational search space, or during scoring, i.e. *a posteriori* by filtering or scoring the generated models based on the discrepancy between the experimentally measured structural properties and the back-calculated ones. In this work, we focus on the latter and assess the applicability of low-resolution shape data, obtained from either IMMS or SAXS experiments, for scoring decoys generated by macromolecular docking using our information-driven docking approach HADDOCK [14,15]. The reasons to focus first on these experimental techniques are that they are rapid, effective, and can be applied to a very broad mass range [1,4,9,10].

IMMS is a combination of two different spectrometry techniques: Ion mobility and mass spectrometry that are working under gas phase conditions. Coupling IM with MS provides information about the mass, subunit composition and Collision Cross Sections (CCS) of the biomolecular complexes under study [2,10,16]. The CCS is defined as the rotationally averaged area of a molecule that is available to interact with the buffer gas, which entails low-resolution (one-dimensional) shape-related information [10]. It has been demonstrated that CCS values estimated for the

lowest charge states measured under gas phase conditions often correlate with the ones simulated from X-Ray and NMR structures [10,17–19]. So far CCS-guided modeling has been conducted by comparing experimental CCS values with the ones simulated from models. It has been used to understand various important biological phenomena, e.g. virus capsid formation [20], aggregation of amyloid fibrils [21], protein folding pathways [22] and potential conformational changes [16]. Lately D'Abramo *et al.* used CCS data to score the complexes generated by docking [23]. They tested their CCS-based scoring function on available CAPRI (Critical Assessment of PRediction of Interactions) targets and observed that the probability of selecting a near-native solution was significantly higher than a random selection.

SAXS allows characterization of biomolecular complexes under native conditions [1,24]. It measures the intensity scattered by a protein sample at low scattering angles. From the scattering profile size- and shape-related information can be extracted, such as molecular weight, radius of gyration, maximum molecular dimension and a low-resolution 3D molecular envelope [1,4,25]. The fact that SAXS experiments can be conducted in a rapid manner has recently made it possible to run them in a high-throughput manner [26]. Over the years, SAXS has been used to shed light on challenging vital biological processes, like folding-unfolding events, conformational changes and oligomerization processes [1,4,24]. To this end, the SAXS information has been typically translated into: (i) *a molecular envelope*, which can guide the construction of *ab intio* bead models or the docking of individual 3D structures [24,25,27], (ii) *a restraining energy term*, which can be used to refine structure directly against the SAXS curve (often in combination with orientational NMR restraints) [4,28,29], or (iii) *a scoring term* that calculates the discrepancy between the experimental scattering curve and the back-calculated one from the model [30–32]. SAXS-based scoring has been recently incorporated into two *ab initio* docking methods, pyDock [33] and PatchDock [34], revealing that SAXS integrated scoring improves the accuracy of model selection significantly.

Here we present a thorough and systematic investigation of the information-content of both CCS and SAXS data and their usefulness in filtering docking decoys based on a benchmark of 176 complexes (Docking Benchmark 4.0) [35]. For this we used HADDOCK, our information-driven docking program that allows inclusion of various types of sparse experimental data to drive the modeling of biomolecular complexes [14,15]. The docking procedure is composed of three stages: Initial docking by rigid-body energy minimization (*it0*), semi-flexible refinement in torsion angle space (*it1*) and final refinement in explicit solvent (*water*). The binding mode of the complex is roughly determined during *it0* and then the top scoring models (typically the top 200-400, ranked according to the HADDOCK score) are selected for further refinement. In this study, we demonstrate that this selection process plays

a critical role in the accuracy of the final models. For this, we combine two different scoring functions, integrating CCS information from IMMS or SAXS scattering profiles, with the conventional HADDOCK score and measure their performance on docking decoys obtained by running HADDOCK in *ab initio* mode. We analyze the strengths and weaknesses of each data type within the context of macromolecular docking and reveal that the highest accuracy in scoring can be obtained when these terms are combined with the conventional HADDOCK score. Last but not least, this study represents HADDOCK's base-line performance, both for *ab initio* docking and for the inclusion of shape data as filter.

## 2. Materials and Methods

### *2.1. Running HADDOCK in ab inito mode*

All complexes were docked using the *ab initio* mode of HADDOCK2.2 with only *center-of-mass* (CM) restraints. CM restraints are distance restraints defined between the centers-of-mass of the molecules. They are automatically calculated from the dimensions of each component along its principal *x*,*y*,*z* axis ($d_x$, $d_y$, $d_z$) as:

$$d_{CM} = \sum_{i=1}^{N\leq 6} (d_{x,i}+d_{y,i}+d_{z,i}-d_{max,i})/4 \quad \text{(Eq. 1)}$$

,where *N* is the number of constituents that are docked, $d_{x,y,z}$ the dimensions along each principal axis and $d_{max}$ the longest dimension (the latter is subtracted to ensure generation of tighter restraints). From this a distance restraint is defined with an upper distance bound set to ($d_{CM}$+*1*) Å and no lower bound.

For each docking, 10,000 rigid-body structures were generated with *ntrials*=5 (meaning that 5 docking trials were performed and the best solution was kept) starting from random orientations. For each solution, the 180° rotated solution was also automatically sampled, resulting in an effective sampling of 100,000 decoys, of which only 10,000 were written to disk. The number of structures for subsequent refinement (*it1*, *water*) was increased to 400. In the case of symmetrical multimers, proper non-crystallographic symmetry (NCS) and symmetry restraints were used as described previously [36]. The solutions were ranked at the end of each docking stage according to the following HADDOCK scores:

$$\textbf{it0:}\ 0.01E_{vdW} + 1.0E_{Elec}+0.01E_{CM} -0.01BSA+1.0E_{Desolv} +0.1E_{Sym} \quad \text{(Eq. 2)}$$

$$\textbf{it1:}\ 1.0E_{vdW} + 1.0E_{Elec}+0.\ 1E_{CM} -0.01BSA+1.0E_{Desolv} +0.1E_{Sym} \quad \text{(Eq. 3)}$$

$$\textbf{water:}\ 1.0E_{vdW} + 0.2E_{Elec}+0.1E_{CM} +1.0E_{Desolv} +0.1E_{Sym} \quad \text{(Eq. 4)}$$

$E_{vdW}$ is the van der Waals intermolecular energy, $E_{Elec}$ the intermolecular electrostatic energy, $E_{CM}$ the distance restraint energy, $E_{Desolv}$ the empirical desolvation energy [37],

BSA the buried surface area and, if present, $E_{Sym}$ the symmetry restraint energy. The non-bonded interactions ($E_{vdW}$ and $E_{Elec}$) are calculated using an 8.5Å cutoff using OPLS parameters [38]. The final models were clustered based on the pairwise ligand interface RMSD with a minimum cluster size of four and a RMSD cut-off of 7.5Å. The resulting clusters were ranked based on the average score of their top four members.

### *2.2. Generation of synthetic CCS values*

The Leeds method [39] was used to generate synthetic CCS values of the native complexes and of the docking models. Leeds is a Monte Carlo approach estimating the area of the protein to which the buffer gas can collide. It is the fastest among the other CCS simulation methods [40–43], and its' predictions have been shown to be in good agreement (~7% difference) with experimental values [10,39]. During simulations the default settings of Leeds were kept with the choice of Helium as the buffer gas. In order to measure the discrepancy between the CCS value of the native complex ($CCS_{Refe}$) and of the model ($CCS_{Mod}$), the following fit term was defined: $CCS_{Fit} = abs(CCS_{Refe} - CCS_{Mod})/CCS_{Refe}$ (Eq. 5)



### *2.3. Generation of synthetic SAXS curves*

A commonly used method, Crysol [44], was applied to simulate the SAXS curves of the native complexes and of the docking models. Synthetic SAXS data were simulated for the momentum transfer (s) range of 0.0005-0.5 with default parameters, except for the maximum order of harmonics and the number of data points generated, which were set to 18 and 256, respectively. In order to measure the impact of SAXS data on docking in a realistic manner, noise was added to the data such that to mimic the experimental error, as in previously published examples [45–47]:

- The *error-to-intensity ratio* ($k_{exp}$) was calculated by using the data points of a good quality experimental SAXS curve (measured from a sample having a concentration >10 mg/ml (Tobias Madl, personal communication)) by:

$$k_{exp}(s) = \sigma_{exp}(s)/I_{exp}(s) \text{ (Eq. 6)}$$

- A second order Gaussian Error Function was fit to the distribution of $k_{exp}$ as a function of the momentum transfer s, in order to simulate the *error-to-intensity ratio* ($k_{sim}$):

$$k_{sim}(s) = 0.22exp\left[\left(-\frac{(s-0.42)}{0.20}\right)^2\right] + 0.06exp\left[\left(-\frac{(s-0.24)}{0.08}\right)^2\right] \quad \text{(Eq. 7)}$$

For a realistic error estimation, $k_{sim}$ was randomly chosen within the confidence interval (95%) of the Gaussian Error Function (by using the *rand* function of MATLAB [48], for details see Supplementary Material). This ratio was then used to scale the intensity predicted by Crysol:

$$\sigma_{sim}(s) = I_{CRYSOL}(s) \cdot k_{sim}(s) \quad \text{(Eq. 8)}$$

To measure the fit between the SAXS curve of the native complex and of the model, the discrepancy value, $\chi$ (calculated by Crysol) was used:

$$\chi^2 = \frac{1}{N_p}\sum_{i=1}^{N_p}\left[\frac{I_e(s_i) - cI(s_i)}{\sigma(s_i)}\right]^2 \quad \text{(Eq. 9)}$$

,where $N_p$ is the total number of experimental data points, $\sigma(s_i)$ is the experimental error, c is a scaling factor, $I_e(s_i)$ is the experimental intensity and $I(s_i)$ is calculated intensity [44]. During fitting, the constant subtraction option was used to improve the fit [44].



### *2.4. Construction of the new scoring function*

The CCS and SAXS fit terms defined above (Eq. 5 and 9) were individually incorporated into the standard HADDOCK score. Their optimum weight was determined by exploring a range between 0-500 and selecting the weight that results in the maximum number of benchmark cases with at least one hit at their top 400. This optimization was performed using the *it0* structures generated for the 176 cases of the Docking Benchmark 4.0 [35] (see below) by running HADDOCK in *ab initio* mode. As a result, the new scoring functions were determined:

$$HADDOCK_{CCS} = HADDOCKScore + 50 \cdot CCS_{Fit} \quad \text{(Eq. 10)}$$
$$HADDOCK_{SAXS} = HADDOCKScore + 50 \cdot \chi \quad \text{(Eq. 11)}$$

In order to define the base-line performance, a random selection of various numbers *X* of models was performed. A near native hit in the top *X* can be obtained if there are $\geq 10000/X$ near native solutions generated for that particular case. For each top *X* category, an enrichment factor was calculated by averaging over all complexes the ratio of the number of hits selected by the combined scoring function (either $HADDOCK_{CCS}$ or $HADDOCK_{SAXS}$) to the number of hits selected by random selection.

### *2.5. Docking Benchmarks*

The first benchmark used in this study is Docking Benchmark 4.0 [35], which consists of 176 unbound-unbound cases, composed of 52 enzyme-inhibitor, 25 antibody-antigen and 99 other type of complexes. As SAXS and IMMS are sensitive to mass change, the number of residues of the unbound states was matched to the one of the native complex (i.e. missing parts in the structure of complex were removed from the structures of the free forms).

The second benchmark is composed of multimeric complexes with more than two components. It is an extension of the previously published benchmark [36,49] and contains nine cases: four homotrimers, two homotetramers and three homopentamers, with C3, D2 and C5 symmetry respectively (**Table 1**). For six of them (1A3F, 1QU9, 1OUS, 1VIM, 1VPN, 1C4Q), the docking was started with the separated components of the crystal structure ("bound docking"), since the unbound coordinates are not available.



**Table 1. The properties of the Symmetric Multimer Benchmark.**

| PDB ID | Complex Type | Number of amino acids (per chain) | Shape of each chain / Anisotropy value |
|---|---|---|---|
| 1JS0[u] [50] | $C_3$ trimer | 124 | Prolate / 2.2 |
| 1QU9[b] [51] | $C_3$ trimer | 128 | Prolate / 0.1 |
| 1A3F[b] [52] | $C_3$ trimer | 137 | Prolate / 1.5 |
| 1URZ[u] [53] | $C_3$ trimer | 400 | Prolate / 19 |
| 1OUS[b] [54] | $D_2$ tetramer | 114 | Prolate / 2.8 |
| 1VIM[b] [55] | $D_2$ tetramer | 200 | Prolate / 0.7 |
| 1B0C[u] [56] | $C_5$ pentamer | 58 | Prolate / 3.6 |
| 1C4Q[b] | $C_5$ pentamer | 69 | Spherical / 0.0 |
| 1VPN[b] [57] | $C_5$ pentamer | 289 | Prolate / 4.4 |

[b] Docking was started with the separated components of the crystal structure.
[u] Docking was started with the free forms of the monomers.

### *2.6. Classification of the cases according to their shape*

We classified the shape anisotropy (SA) of the various systems in order to study the shape-dependency of the docking and scoring results. Since, in a real case, the structure of the complex will not be known, this classification was performed on the largest component of each complex. If the partners were similar in size, the one

having the largest SA was considered. The shape anisotropy was calculated from the eigenvalues ($\lambda_i$) of the molecule's *Gyration Tensor* (*G*) [58,59]:

$$SA = \frac{\prod_{i=1}^{3}(\lambda_i - \bar{\lambda})}{\bar{\lambda}^3} \quad (12)$$

,where $\bar{\lambda}$ is the mean of the eigenvalues. For a rod-like *prolate* protein, *SA* would be larger then zero (as $\lambda_1 \gg \lambda_2 \approx \lambda_3$); for a disc-like *oblate* protein it would be smaller than zero (as $\lambda_1 \ll \lambda_2 \approx \lambda_3$) and for an isotropic *sphere*-like protein *SA* would be equal to zero (as $\lambda_1 \approx \lambda_2 \approx \lambda_3$).

### *2.7. Assessment of the docking models quality*

At the rigid-body docking stage, a solution with i-RMSD ≤ 4.0 Å was considered to be a *hit*; a case containing at least one hit within the top 400 was considered to be *successful*. After *water refinement*, the models were evaluated based on the CAPRI criteria [60]:

- Acceptable prediction (one star): i-RMSD ≤ 4 Å or l-RMSD ≤ 10 Å
- Good prediction (two stars): i- RMSD ≤ 2 Å or l-RMSD ≤ 5 Å
- High-quality prediction (three stars): i- RMSD ≤ 1 Å or l-RMSD ≤ 1 Å

over the backbone atoms of the ligand (rigid component) after fitting on the receptor. A cluster was considered of one-, two-, or three-star quality if at least one of its top four members was of the corresponding quality.

## 3. Results and Discussion

### *3.1. The Performance of HADDOCK ab initio is mainly limited by scoring rather than sampling*

HADDOCK was run in *ab initio* mode to dock the unbound structures of the Docking Benchmark 4.0 (see **Sections 2.1.** and **2.5.**). At the rigid-body docking stage at least one hit (near-native model among the 10,000 rigid-body docking models) was obtained for 78% of the whole benchmark, corresponding to 138 cases out of 176. Within those 138 *successful* cases, at least one near-native model could be ranked in the top 400 for 49%, top 100 for 27% and top 10 for 13% of the cases. These statistics reveal a large success rate difference between sampling and scoring: considering the 138 *successful* cases, even for the best case (top 400 category), the overall success rate dropped upon scoring by 50% (from 100% to 49%). This indicates that the success rate of *HADDOCK ab initio* is mainly limited by scoring rather than sampling [61]. A possible solution to overcome this problem is to incorporate external information into the scoring function [6,62–65]. For this purpose

we developed two different scoring functions that integrate collision cross section information from IMMS and scattering profile from SAXS experiments.

### *3.2. CCS does not discriminate between different docking poses of dimers*

In order to evaluate the effect of CCS on HADDOCK's scoring function, the set of 138 *successful* complexes with at least one near native hit were ranked according to a random selection, the standard HADDOCK score (Eq. 2), $CCS_{Fit}$ only (Eq. 5) and the optimized combined $HADDOCK_{CCS}$ score (Eq. 10). The corresponding scoring performances is depicted in **Figure 1A**. This analysis revealed that the success rate of $CCS_{Fit}$-based ranking was either significantly worse than the random selection (as in top 400 category) or similar to it. This implies that for this particular benchmark set the CCS data do not contain any information that would allow discriminating different docking poses. Accordingly, incorporation of the $CCS_{Fit}$ term into the standard HADDOCK score did not result in any improvement.

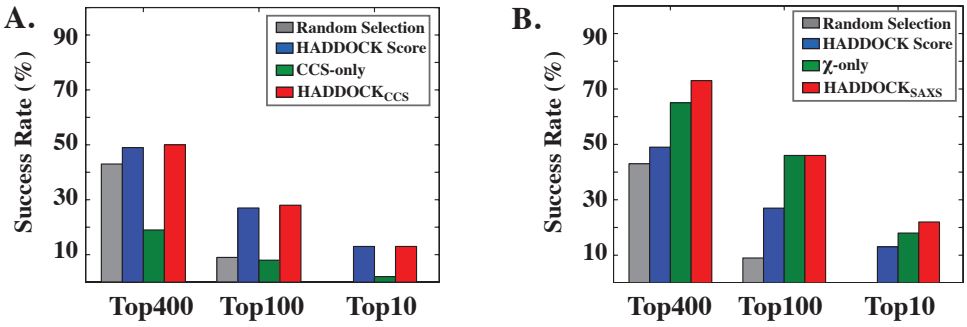



**Figure 1. Depicted success rates of various scoring functions.** The success rates were calculated on Docking Benchmark 4.0 cases, for which HADDOCK *ab inito* could generate at least one hit (138 cases). For each top ranking category, the scoring performances are presented in the order of **A.** random selection, HADDOCK, $CCS_{Fit}$-only and $HADDOCK_{CCS}$ scores, and **B.** random selection, HADDOCK, χ-only and $HADDOCK_{SAXS}$ scores.

Previously Ruotolo *et al.* presented that simulated CCS values of a complex, composed of 8kDa subunits (~75 amino acids) start to diverge from each other significantly only at large subunit numbers (varying between 8-to-12 subunits, see **Figure 1a** of Ruotolo *et al.* [2]). This observation was further supported by the findings of Politis *et al.* (**Figure 3a** of [16]) and Pukala *et al.* (**Figure 2** [66]). In our case, the lack of sensitivity of the $CCS_{Fit}$ term could be attributed to the fact that the Docking Benchmark 4.0 consists only of dimers or rather limited size (between 7-90 kDa). To investigate the subunit- and mass-dependency of CCS, we scored the *it0* models of 1VIM and 1VPN, the two largest complexes of our multimer benchmark,

with $CCS_{Fit}$ and $HADDOCK_{CCS}$ terms. In the case of 1VIM, both $CCS_{Fit}$ and $HADDOCK_{CCS}$ could rank all of the near-native solutions at top 10, whereas for 1VPN they could not rank any hit even at top 400. Although these results were not conclusive, they suggest that CCS data do carry potential for the study of large macromolecular complexes.

### *3.3. $HADDOCK_{SAXS}$ improves scoring of rigid-body docking models*

In order to assess the scoring ability of $HADDOCK_{SAXS}$, the success rates of a random selection, the standard HADDOCK score, the SAXS only $\chi$-score and the combined $HADDOCK_{SAXS}$ score were compared (**Figure 1B**). The success rate statistics revealed that the SAXS $\chi$ has a significant discriminative ability (even higher than the standard HADDOCK score). Among all the scoring functions, however, the combination of $\chi$ and HADDOCK score performed the best, especially for the top 400 category (**Figure 1B**), with a 24% improvement in success rate (from 49% to 73%, measured over the set of 138 complexes with at least one near-native hit). We also calculated the enrichment factor of each scoring function compared to a random selection (**Table 2**, see **Section 2.4.**). On average $HADDOCK_{SAXS}$ scoring could enrich the number of hits selected per case by 8 fold for the top 400 category and by 17 fold for the top 100 category, showing the best performance and almost doubling the performance of the HADDOCK score.



**Table 2. Enrichment factors of various scoring functions compared to a random selection.** The enrichment factor is calculated as the ratio of the number of selected near-native solutions (for the given top category) compared to a random selection of the same number of models. (This measure could only be calculated for top 400 and top 100 categories, as for the lowers ones random selection could not rank any hit in the given top category.)

| | **Haddock** | **$\chi$-only** | **$HADDOCK_{SAXS}$** |
|---|---|---|---|
| **Top 400** | 4.5 | 5.7 | 7.9 |
| **Top 100** | 9.4 | 14.5 | 16.8 |

In the last stage of HADDOCK, the water-refined solutions are typically clustered requiring a minimum number of four models per clusters. Successful clustering would therefore require a minimum number of four near-native solutions at *it0*. Considering this requirement, only 116 out of the 176 complexes would qualify for clustering (irrespective of the rank of the models). While the resulting overall success rate decreases from 78% to 66%, the overall conclusions about the improved performance of the combined SAXS score do not change (data not shown).

It has already been demonstrated that SAXS could distinguish better between different conformations of the same complex, if the anisotropy of the complex's constituents is high, namely if their shape is far from spherical [4,33,67]. To explore this point we classified the *successful* set of 138 complexes into three: *prolate* (90 cases), *oblate* (25 cases), *spherical* (23 cases) based on the largest constituent of each complex and not the complex itself, so that this measure can potentially be used as a predictor of the impact of SAXS scoring in real cases, where the structure of the complex is still unknown. The performance of the various scoring functions was analyzed for each class separately revealing that SAXS χ-based scoring performs significantly better for anisotropic (66% for *prolate*, 72% for *oblate*) complexes than for *spherical* (57%) ones (**Figure 2**). The same can be observed for the combined HADDOCK$_{SAXS}$ score, although with an increased overall performance (74 % for *prolate*, 76% for *oblate*, 65 % for *spherical*).

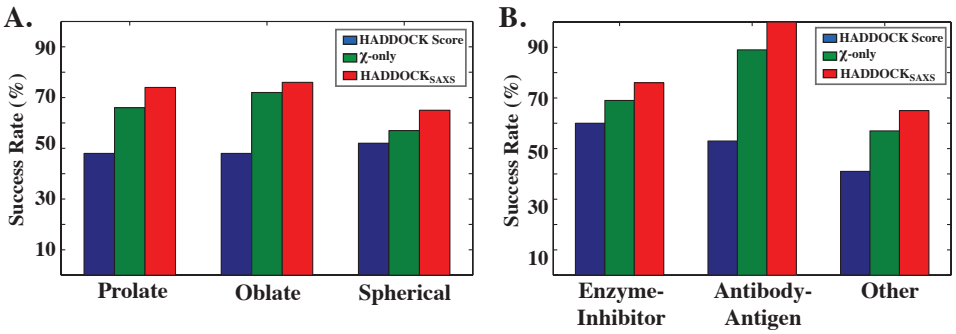



**Figure 2. Depicted success rates of various scoring functions classified according to shape and function of the complexes.** The success rates presented here are for top 400 category and they are calculated on 138 *successful* cases of the benchmark. **A.** The *successful* docking decoys are classified according to their shape determinants into: *prolate* (90 cases), *oblate* (25 cases), *spherical* (23 cases), and **B.** biochemical functions into: Enzyme-Inhibitor (45 cases), Antibody-Antigen (19 cases), Other (74 cases)**.**

When the same set was classified according to biochemical function, the most pronounced increase in scoring accuracy was seen for Antibody-Antigen complexes: HADDOCK$_{SAXS}$ could rank at least one hit at top 400 for all of the Antibody-Antigen cases (100% success rate), whereas the regular HADDOCK score could do the same only for half of them. This impressive improvement is related to the shape dependency of SAXS scoring since all of the Antibody-Antigens complexes are highly anisotropic (*prolate*). Compared to Antibody-Antigens complexes, the success rate of HADDOCK$_{SAXS}$ dropped to 76% for Enzyme-Inhibitors and 65% for the Other category.

### *3.4 Impact of flexible refinement*

The top 400 models re-ranked according to the $HADDOCK_{SAXS}$ score were subjected to HADDOCK's two-step refinement protocol (*it1, water*). This flexible refinement resulted in a success rate increase of 2% compared to rigid-body docking (73%). Among the cases having at least one hit, 56% contained one-star quality solutions, 42% two- and only 2% three-star solutions. The performance comparison between HADDOCK and $HADDOCK_{SAXS}$ revealed that $HADDOCK_{SAXS}$ improves the success rate of the high-ranking top categories (**Figure 3**).

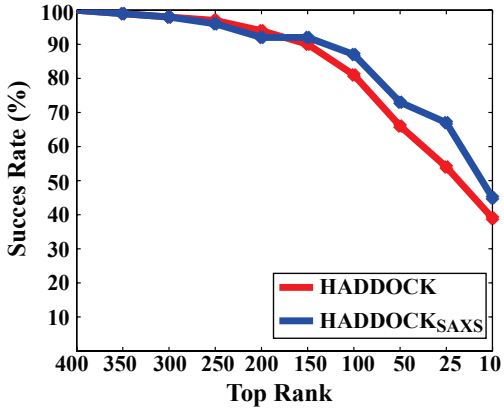



**Figure 3. Performance comparison of HADDOCK (blue) and $HADDOCK_{SAXS}$ (red) scores at the end of the *water* stage.** The success rates were calculated for different top ranking categories (between top 400 and top 10) of Docking Benchmark 4.0 cases, for which HADDOCK *ab inito* could generate at least one hit (138 cases).

For example, for the top 25 category $HADDOCK_{SAXS}$ could increase the success rate of HADDOCK by 17% (from 54% to 71%). Note again that these numbers refer to the scoring performance of HADDOCK run in *ab initio* mode, without any additional information except for the SAXS term used in scoring. As an alternative to HADDOCK *ab inito*, docking can also be driven by bioinformatic predictions. In that case, the scoring performance was shown to double, from 27% to 58%, after water refinement [61].

### *3.5. $HADDOCK_{SAXS}$ improves the ranking of symmetric multimers especially for anisotropic ones*

In order to assess the impact of SAXS scoring on larger systems, we run HADDOCK in *ab initio* mode on a symmetric multimer benchmark composed of 9 complexes from trimers to pentamers (**Table 1**). This benchmark is an extension of our previously reported multibody docking work [36]. HADDOCK generated a substantial number of near-native hits for five cases (**Table 3**). Their number however decreased significantly for larger system sizes (as in 1OUS, 1VIM, 1VPN, **Table 3**) or in cases where significant conformational changes are observed upon binding, such

as in 1URZ where backbone-RMSD change is 4.3Å. For these, very few near-native solutions could be sampled or none as in the latter case.

**Table 3. The rigid-body docking (*it0*) performance of HADDOCK and scoring performance of HADDOCK$_{SAXS}$ on Symmetric Multimer Benchmark.** The second column refer to the total number of near-native hits[a] generated, the third to the number of near native hits ranked within the top 400 using the regular HADDOCK score (Eq. 4) and the fourth column to the number of hits ranked using the combined HADDOCK$_{SAXS}$ score (Eq. 11). The last column corresponds to the enrichment factor calculated as the ratio of the number of hits by combined HADDOCK$_{SAXS}$ scoring to the number of hits by regular HADDOCK scoring.

| PDB ID | Number of hits generated | Number of hits ranked by HADDOCK (% of total hits) | Number of hits ranked by HADDOCK$_{SAXS}$ (% of total hits) | Enrichment factor compared to the standard HADDOCK score |
|---|---|---|---|---|
| 1JS0 | 90 | 4 (4%) | 88 (98%) | 24.5 |
| 1QU9 | 409 | 368 (92%) | 366 (91%) | 1.0 |
| 1A3F | 418 | 112 (28%) | 273 (68%) | 2.4 |
| 1URZ | - | - | - | - |
| 1OUS | 1 | 0 | 0 | 0 |
| 1VIM | 8 | 8 (100%) | 8 (100%) | 1 |
| 1B0C | 66 | 0 | 30 (45%) | ∞ |
| 1C4Q | 231 | 217 (94%) | 212 (92%) | 1.4 |
| 1VPN | 8 | 5 (62%) | 5 (62%) | 1 |

[a]A solution is considered to be a near-native hit, if it is within 4.0Å i-RMSD or 10Å l-RMSD of the native solution (see **Section 2.7.**).

Irrespective of the docking difficulty, HADDOCK$_{SAXS}$ could rank at least one hit within the top 400 for seven out of eight cases, which translates into a success rate of 88%. As observed before for dimers, also in the case of multimers it performs significantly better for systems having high anisotropy (1A3F, 1JS0, 1B0C) than for the ones with low anisotropy (1QU9, 1C4Q) (**Table 1**). At the end of water refinement, a top ranking high accuracy three-star solution could be generated for four cases, a medium quality two-star for two and an acceptable one-star solution for one (**Table 4**, **Figure 4**). Moreover, all these high quality solutions were populated in top ranking clusters.

**Table 4. Scoring performance of HADDOCK$_{SAXS}$ on Symmetric Multimer Benchmark at the end of the flexible refinement (*water*) stage.**

| | Single structure scoring (Quality[a]/Rank) | Cluster-based scoring (Quality[a]/Rank) |
|---|---|---|
| 1JS0 | ★/1 & ★★/14 | ★/1 & ★★/2 |
| 1QU9 | ★★★/1 | ★★★/1 |
| 1A3F | ★★★/1 | ★★★/1 |
| 1URZ | - | - |
| 1OUS | - | - |
| 1VIM | ★/1 | ★/1 |
| 1B0C | ★/7 | ★/1 & ★★/2 |
| 1C4Q | ★★★/1 | ★★★/1 |
| 1VPN | ★★★/1 | not clustered |

[a] For the definition of quality, see **Section 2.7**.



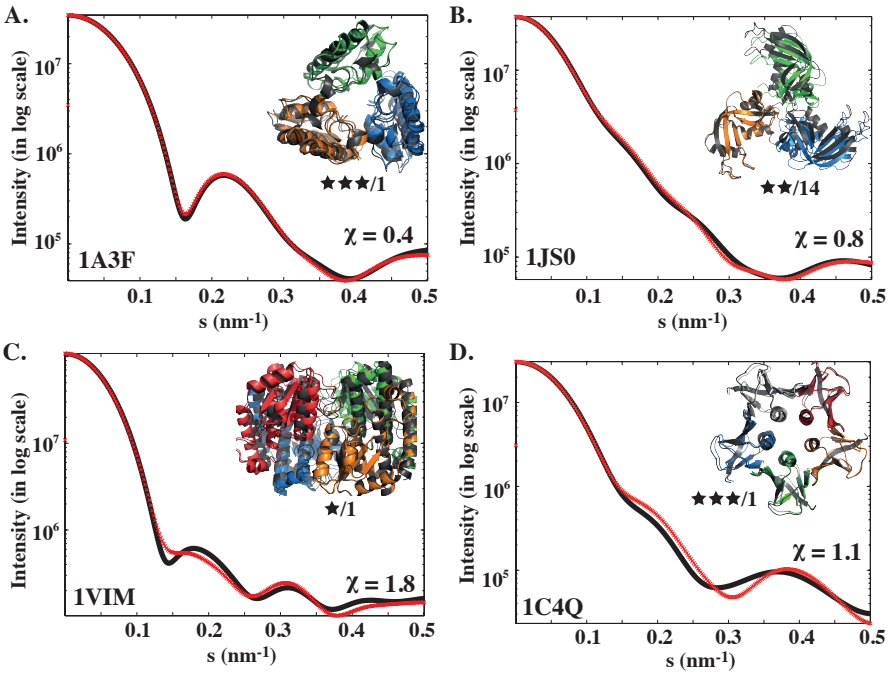



**Figure 4. Selected illustrations of the best models ranked by HADDOCK$_{SAXS}$.** For each case, the fit between the simulated scattering curve of the crystal structure (red crosses) and the one of the best HADDOCK model (black, connected-lines) is demonstrated. The structure of the best HADDOCK model (colored) superimposed on top of the crystal structure (black) is placed at the top right corner of each scattering curve. Next to each model, it's quality and rank are indicated.

### *3.6. Application examples with experimental SAXS profiles*

We tested the performance of HADDOCK$_{SAXS}$ on three cases (3K3K, 2R15 and 1O6S) for which experimental SAXS profiles are available (**Table 5**). The experimental SAXS curve of 3K3K was acquired from the Biosis database [26], the profiles of the other two were kindly provided by Dmitri Svergun [33]. For docking, HADDOCK *ab initio* and for scoring, HADDOCK$_{SAXS}$ were used in the same way as described above. The data and results are summarized in **Tables 5** and **6**.

3K3K is an abscisic acid receptor forming an asymmetric homodimer [68]. Docking of 3K3K was started from the isolated chains of 3K3K. At the end of *it0*, HADDOCK could rank 63% of the hits generated within the top 400, whereas this rate dropped to 44% upon HADDOCK$_{SAXS}$ scoring (**Table 6A**). This accuracy drop was due to the fact that the $\chi$ value calculated for a degenerate binding mode and for the native solution was the same (**Figure 5A,B**). Recently Gabel *et al*. [67] remarked that, when the monomers of a homodimer are oriented in a "*side-by-side*" fashion, as in the case here, the SAXS curves simulated from different *side-by-side* orientations of the same complex fit equally well to the experimental curve. This degeneracy issue is less problematic when the monomers are arranged in an "*end-to-end*" (rod-like) fashion. Despite the degeneracy problem, the contribution of the HADDOCK score ensured that a biologically meaningful, high accuracy cluster was ranked at top at the end of the flexible refinement (**Figure 5B**).



**Table 5. Test benchmark with experimental SAXS curves.**

| PDB ID | Complex Type | # of amino acids (per chain) | Shape of each chain / Anisotropy value |
|---|---|---|---|
| 3K3K[b] [68] | Homodimer | 211 | Prolate / 0.8 |
| 2R15[b] [69] | Homodimer | 212 | Prolate / 26.8 |
| 1O6S[u] [70] | Dimer | 466/105 | Prolate / 11.3 |

[b] Docking was started with the separated components of the crystal structure.
[u] The docking was started with the free forms of the monomers: 1O6T [70] and chain B of 1FF5 [71].

2R15 is a complex of an end-to-end myomesin fragment, composed of two highly *prolate* domains (My12 and My13) (**Table 5**) [69]. My12 was used as starting structure for both chains. During rigid-body docking, only two near-native models could be generated, which were successfully selected by HADDOCK$_{SAXS}$ for further refinement. At the end of the flexible refinement one of these near-native solutions was ranked within the top 10 (**Table 6B**). This scoring success (successful selection of

2 out 10,000 models) could be achieved since the monomers of 2R15 are highly *prolate* and, in the complex, they are oriented in an "*end-to-end*" fashion (**Figure 5C**).

The last example, 1O6S is a *prolate* complex of internalin (InlA) and the N-terminal domain of human E-cadherin (hEC1) (**Table 5**) [70]. The docking was started with the free forms of In1A (PDB id, 1O6T [70]) and hEC1 (PDB Id, 1FF5_chainB [71]). During_rigid-body docking 25 near-native models were generated, none of which were ranked within the top 400 with the regular HADDOCK score. Yet $HADDOCK_{SAXS}$ selected 80% of those near-native models for further refinement, which were then populated in the top ranking one-star quality cluster at the end of the flexible refinement (**Table 6**, **Figure 5D**).

**Table 6A. Rigid-body docking (*it0*) performance of HADDOCK and scoring performance of $HADDOCK_{SAXS}$ on the Benchmark with experimental SAXS curves.** The second column refer to the total number of hits generated, the third column to the number of hits ranked at top 400 by HADDOCK and the last column to the number of hits ranked by $HADDOCK_{SAXS}$.

| | # of hits generated | # of hits ranked by HADDOCK (% of total hits) | # of hits ranked by $HADDOCK_{SAXS}$ (% of total hits) |
|---|---|---|---|
| 3K3K[b] | 117 | 74 (63%) | 52 (44%) |
| 2R15 | 2 | 0 (%) | 2 (100%) |
| 1O6S | 25 | 0 (%) | 20 (80%) |

**Table 6B. Scoring performance of $HADDOCK_{SAXS}$ at the end of flexible refinement (*water*) stage.**

| | Single structure scoring (Quality[a]/Rank) | Cluster-based scoring (Quality[a]/Rank) |
|---|---|---|
| 3K3K[b] | ★★★/17 | ★★★/1 |
| 2R15 | ★/6 & ★★12 | - |
| 1O6S | ★/32 | ★/1 |

[a] For the definition of quality, see **Section 2.7**.

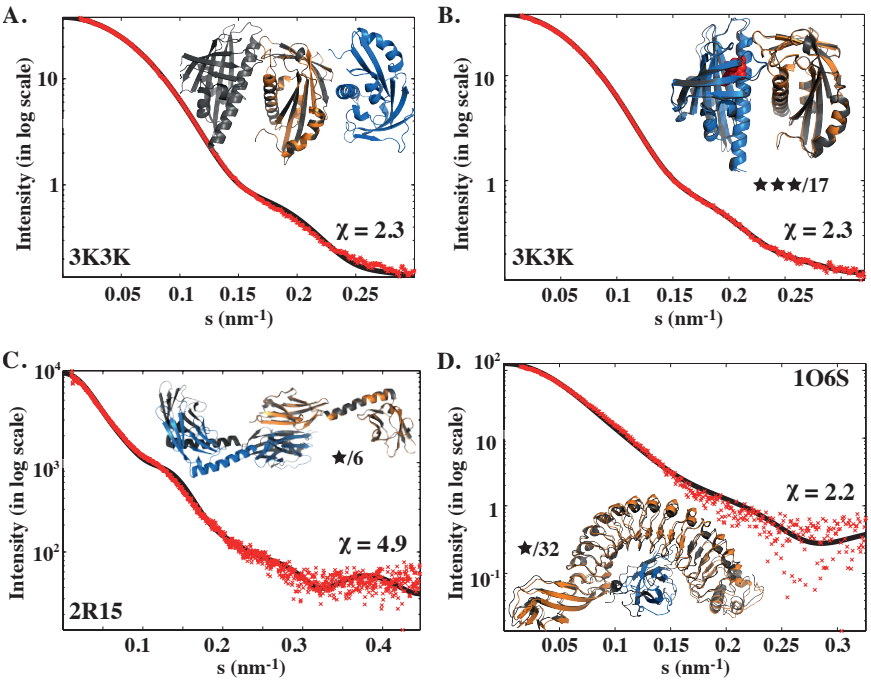



**Figure 5. The performance of HADDOCK$_{SAXS}$ on cases with real SAXS profile.** For each case, the fit between the experimental scattering curve (red crosses) and the simulated one of the HADDOCK model (black, connected-lines) is depicted together with the superimposed HADDOCK model (colored) on top of the crystal structure (black). Next to each model, it's quality and rank are indicated. **A-B.** When the monomers of a homodimer are oriented in a "*side-by-side*" fashion, as in the 3K3K case, SAXS curves simulated from **A.** a degenerate and **B.** a correct binding mode could fit equally well to reference SAXS curve. **C.** For 2R15, although two near-native models were sampled during *it0*, one of them could be ranked at top 10 with HADDOCK$_{SAXS}$ at the end of *water* refinement. **D.** HADDOCK$_{SAXS}$ could select 80% of the near-native 1O6S models for further refinement, which were then populated in a top ranking one-star quality cluster at the end of *water*.

## 4. Conclusions

In this work, we demonstrated that, in the absence of external information, the accuracy of docking models obtained by running HADDOCK in *ab initio* mode is substantially limited by scoring rather than sampling. The inaccurate nature of conventional scoring functions can, however, be improved by incorporating information-based terms. Within this context, we evaluated the impact of integrating CCS from IMMS and scattering profile from SAXS experiments into HADDOCK's scoring function.

Our analysis revealed that the information content of CCS data is of too low-resolution for distinguishing correct models in a large docking decoy set composed of dimers. Despite this failure to distinguish near-native docking poses for dimeric assemblies, in principle, CCS data could add useful information for large macromolecular complexes. Indeed, various literature examples have underpinned the size dependency of CCS, indicating that CCS-based terms start to be discriminative at high subunit numbers [2,16,66]. In order to test this, we re-ranked the rigid-body docking models of the two largest assemblies of our multimer benchmark, 1VIM and 1VPN (**Table 1**). As a result, $CCS_{Fit}$ and $HADDOCK_{CCS}$ worked out nicely for 1VIM (both of them ranked all of the near-native solutions at top), whereas they could not select any hit in the latter for further refinement. This analysis suggests that there is room for further improvements, especially for large assemblies.

In contrast to the CCS data, the results presented here for SAXS scoring show a significant performance improvement, especially when combined with an energy-based function, like the classical HADDOCK score. The scoring performance, however, depends on the shape and arrangement of the complex. Therefore, the SAXS scoring term should be preferentially combined with conventional, energy-based scoring functions, in order to ensure a selection of physically relevant interfaces. Any additional source of information that can be included in the sampling phase will also increase the coverage of near-native solutions and compensate for the inherent shape- and orientation-dependency of SAXS data. For that HADDOCK is an excellent candidate, with its built-in ability of incorporating various sources of data addressing spatial properties, such as interface contacts, orientation between molecules and residue-to-residue distances. The $HADDOCK_{CCS}$ and $HADDOCK_{SAXS}$ scoring functions and associated scripts will be part of a future release of HADDOCK and are freely available upon request.

## Supplementary Material

### *Generation of errors for the synthetic SAXS curve*

The *error-to-intensity ratio* ($k_{exp}$) was calculated by using the data points of a good quality experimental SAXS curve (measured from a sample having a concentration >10 mg/ml (Tobias Madl, personal communication)) (**Figure S1**). A second order Gaussian Error Function provided the best fit to $k_{exp}$ (with $R^2$: 0.97 and RMSE: 0.01), which is used to simulate *error-to-intensity ratio*, $k_{sim}$ for the s-range between 0.02 and 0.5, for the intensity range between -0.5 and 0.25:

$$k_{sim}(s) = a_1 exp\left[\left(-\frac{(s-b_1)}{c_1}\right)^2\right] + a_2 exp\left[\left(-\frac{(s-b_2)}{c_2}\right)^2\right] \quad (1)$$

The coefficients of $k_{sim}$ and their confidence bounds (with 95% of confidence level) were determined to be:

| $a_1$ | $b_1$ | $c_1$ |
|---|---|---|
| 0.2177<br>(0.2158, 0.2196) | 0.4238<br>(0.419, 0.4286) | 0.2044<br>(0.1984, 0.2104) |
| $a_2$ | $b_2$ | $c_2$ |
| 0.0642<br>(0.0574, 0.071) | 0.2407<br>(0.238, 0.2437) | 0.0861<br>(0.0794, 0.09285) |

For a realistic error estimation, $k_{sim}$ was randomly chosen within the confidence interval (95%) of the Gaussian Error Function (by using the *rand* function of MATLAB [48]). An illustration of the randomly simulated $k_{sim}$ is presented in **Figure S1**.

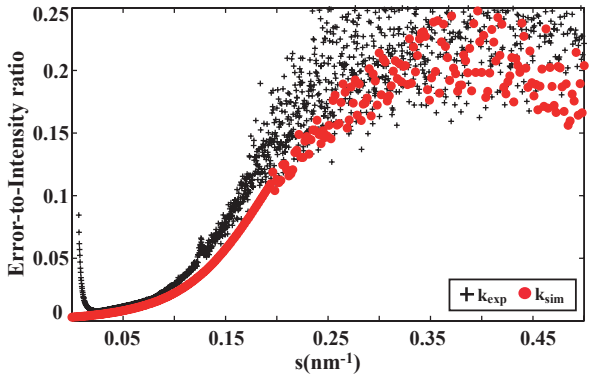



**Figure S1. The evaluation of the real ($k_{exp}$) and simulated ($k_{sim}$) error-to-intensity ratio.** The $k_{exp}$ (black crosses) was calculated by using the data points of a good quality experimental SAXS curve, whereas $k_{sim}$ (red dots) was randomly chosen within the confidence interval (95%) of the Gaussian Error Function (Eq. 1), which was fit to $k_{exp.}$

## References

[1] C.D. Putnam, M. Hammel, G.L. Hura, J.A. Tainer, X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution, Quarterly Reviews of Biophysics. 40 (2007) 191–285.
[2] B.T. Ruotolo, J.L. Benesch, A.M. Sandercock, S.J. Hyung, C.V. Robinson, Ion mobility-mass spectrometry analysis of large protein complexes, Nature Protocols. 3 (2008) 1139–1152.
[3] A.S.J. Melquiond, A.M.J.J. Bonvin, Data-driven docking: using external information to spark the biomolecular rendez-vous, in: Protein-Protein Complexes: Analysis, Modeling and Drug Design, Imperial College Press, 2010: pp. 183–209.
[4] T. Madl, F. Gabel, M. Sattler, NMR and small-angle scattering-based structural analysis of protein complexes in solution, Journal of Structural Biology. 173 (2011) 472–482.
[5] X. Wang, H.-W. Lee, Y. Liu, J.H. Prestegard, Structural NMR of protein oligomers using hybrid methods., Journal of Structural Biology. 173 (2011) 515–29.
[6] F. Alber, F. Förster, D. Korkin, M. Topf, A. Sali, Integrating diverse data for structure determination of macromolecular assemblies., Annual Review of Biochemistry. 77 (2008) 443–477.
[7] G.C. Lander, H.R. Saibil, E. Nogales, Go hybrid: EM, crystallography, and beyond., Current Opinion in Structural Biology. Epub ahead (2012).
[8] J. Rappsilber, The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes., Journal of Structural Biology. 173 (2011) 530–40.
[9] E. Jurneczko, P.E. Barran, How useful is ion mobility mass spectrometry for structural biology? The relationship between protein crystal structures and their collision cross sections in the gas phase, The Analyst. 136 (2011) 20–28.
[10] C. Uetrecht, R.J. Rose, E. van Duijn, K. Lorenzen, A.J. Heck, Ion mobility mass spectrometry of proteins and protein assemblies, Chemical Society Reviews. 39 (2010) 1633–1655.
[11] H.J. Steinhoff, Inter- and intra-molecular distances determined by EPR spectroscopy and site-directed spin labeling reveal protein-protein and protein-oligonucleotide interaction., Biological Chemistry. 385 (2004) 913–20.
[12] A.T. Brunger, P. Strop, M. Vrljic, S. Chu, K.R. Weninger, Three-dimensional molecular modeling with single molecule FRET., Journal of Structural Biology. 173 (2011) 497–505.
[13] J. Garcia-Garcia, J. Bonet, E. Guney, O. Fornes, J. Planas, B. Oliva, Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details, Molecular Informatics. 31 (2012) 342–362.
[14] C. Dominguez, R. Boelens, A.M.J.J. Bonvin, HADDOCK: a protein-protein docking approach based on biochemical or biophysical information., Journal of the American Chemical Society. 125 (2003) 1731–7.
[15] S.J. de Vries, M. van Dijk, A.M.J.J. Bonvin, The HADDOCK web server for data-driven biomolecular docking, Nat Protoc. 5 (2010) 883–897.
[16] A. Politis, A.Y. Park, S.J. Hyung, D. Barsky, B.T. Ruotolo, C.V. Robinson, Integrating ion mobility mass spectrometry with molecular modelling to determine the architecture of multiprotein complexes, PloS One. 5 (2010) e12080.
[17] C.A. Scarff, K. Thalassinos, G.R. Hilton, J.H. Scrivens, Travelling wave ion mobility mass spectrometry studies of protein structure: biological significance and comparison with X-ray crystallography and nuclear magnetic resonance spectroscopy measurements, Rapid Communications in Mass Spectrometry. 22 (2008) 3297–3304.
[18] M. Zhou, C.V. Robinson, When proteomics meets structural biology, Trends in Biochemical Sciences. 35 (2010) 522–529.

[19] J.L. Benesch, B.T. Ruotolo, Mass spectrometry: come of age for structural and dynamical biology, Current Opinion in Structural Biology. 21 (2011) 641–649.
[20] C. Uetrecht, I.M. Barbu, G.K. Shoemaker, E. van Duijn, A.J. Heck, Interrogating viral capsid assembly with ion mobility–mass spectrometry, Nature Chemistry. 3 (2010) 126–132.
[21] S.L. Bernstein, N.F. Dupuis, N.D. Lazo, T. Wyttenbach, M.M. Condron, G. Bitan, et al., Amyloid-β protein oligomerization and the importance of tetramers and dodecamers in the aetiology of Alzheimer's disease, Nature Chemistry. 1 (2009) 326–331.
[22] B.T. Ruotolo, S.J. Hyung, P.M. Robinson, K. Giles, R.H. Bateman, C.V. Robinson, Ion mobility-mass spectrometry reveals long-lived, unfolded intermediates in the dissociation of protein complexes, Angewandte Chemie (International Ed in English). 46 (2007) 8001–8004.
[23] M. D'Abramo, T. Meyer, P. Bernadó, C. Pons, J. Fernández Recio, M. Orozco, On the Use of low-resolution Data to Improve Structure Prediction of Proteins and Protein Complexes, Journal of Chemical Theory and Computation. 5 (2009) 3129–3137.
[24] H.D. Mertens, D.I. Svergun, Structural characterization of proteins and complexes using small-angle X-ray solution scattering, Journal of Structural Biology. 172 (2010) 128–141.
[25] R.M. Buey, P. Chacón, J.M. Andreu, J. Fernando Díaz, Protein Shape and Assembly Studied with X-Ray Solution Scattering: Fundaments and Practice, in: Applications of Synchrotron Light to Scattering and Diffraction in Materials and Life Sciences, Springer Berlin Heidelberg, 2009: pp. 245–263.
[26] G.L. Hura, A.L. Menon, M. Hammel, R.P. Rambo, F.L. Poole, S.E. Tsutakawa, et al., Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS), Nature Methods. 6 (2009) 606–612.
[27] D. Svergun, M.V. Petoukhov, M.H. Koch, Determination of domain structure of proteins from X-ray solution scattering, Biophysical Journal. 80 (2001) 2946–2953.
[28] F. Förster, B. Webb, K.A. Krukenberg, H. Tsuruta, D.A. Agard, A. Sali, et al., Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies., Journal of Molecular Biology. 382 (2008) 1089–106.
[29] A. Grishaev, J. Wu, J. Trewhella, A. Bax, Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data., Journal of the American Chemical Society. 127 (2005) 16621–8.
[30] S. Covaceuszach, A. Cassetta, P.V. Konarev, S. Gonfloni, R. Rudolph, D.I. Svergun, et al., Dissecting NGF interactions with TrkA and p75 receptors by structural and functional studies of an anti-NGF neutralizing antibody, Journal of Molecular Biology. 381 (2008) 881–896.
[31] W. Filgueira de Azevedo, G.C. dos Santos, D.M. dos Santos, J.R. Olivieri, F. Canduri, R.G. Silva, et al., Docking and small angle X-ray scattering studies of purine nucleoside phosphorylase, Biochemical and Biophysical Research Communications. 309 (2003) 923–928.
[32] H. Sondermann, B. Nagar, D. Bar-Sagi, J. Kuriyan, Computational docking and solution x-ray scattering predict a membrane-interacting role for the histone domain of the Ras activator son of sevenless, Proceedings of the National Academy of Sciences of the United States of America. 102 (2005) 16632–16637.
[33] C. Pons, M. D'Abramo, D.I. Svergun, M. Orozco, P. Bernadó, J. Fernández-Recio, Structural characterization of protein-protein complexes by integrating computational docking with small-angle scattering data, Journal of Molecular Biology. 403 (2010) 217–230.
[34] D. Schneidman-Duhovny, S.J. Kim, A. Sali, Integrative structural modeling with small angle Xray scattering profiles, BMC Structural Biology. 12 (2012) 17.
[35] H. Hwang, T. Vreven, J. Janin, Z. Weng, Protein-protein docking benchmark version 4.0, Proteins. 78 (2010) 3111–3114.
[36] E. Karaca, A.S.J. Melquiond, S.J. de Vries, P.L. Kastritis, A.M.J.J. Bonvin, Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server, Molecular & Cellular Proteomics. 9 (2010) 1784–1794.

[37] J. Fernández-Recio, M. Totrov, R. Abagyan, Identification of protein-protein interaction sites from docking energy landscapes., Journal of Molecular Biology. 335 (2004) 843–65.
[38] W.L. Jorgensen, J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, Journal of the American Chemical Society. 110 (1988) 1657–1666.
[39] D.P. Smith, T.W. Knapman, I. Campuzano, R.W. Malham, J.T. Berryman, S.E. Radford, et al., Deciphering drift time measurements from travelling wave ion mobility spectrometry-mass spectrometry studies, European Journal of Mass Spectrometry (Chichester, Eng). 15 (2008) 113–130.
[40] T. Wyttenbach, G. Helden, J.J. Batka, D. Carlat, M.T. Bowers, Effect of the long-range potential on ion mobility measurements, Journal of the American Society for Mass Spectrometry. 8 (1997) 275–282.
[41] T. Wyttenbach, M. Witt, M.T. Bowers, On the Stability of Amino Acid Zwitterions in the Gas Phase: The Influence of Derivatization, Proton Affinity, and Alkali Ion Addition, Journal of the American Chemical Society. 122 (2000) 3458–3464.
[42] A.A. Shvartsburg, M.F. Jarrold, An exact hard-spheres scattering model for the mobilities of polyatomic ions, Chemical Physics Letters. 261 (1996) 86–91.
[43] A.A. Shvartsburg, R.R. Hudgins, P. Dugourd, M.F. Jarrold, P. Dugourd, M.F. Jarrold, Structural information from ion mobility measurements: applications to semiconductor clusters, Chemical Society Reviews. 30 (2001) 26–35.
[44] D. Svergun, C. Barberato, M.H.J. Koch, CRYSOL-a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates, Journal of Applied Crystallography. 28 (1995) 768–773.
[45] P. Bernadó, Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering, European Biophysics Journal : EBJ. 39 (2010) 769–780.
[46] J. Blobel, P. Bernadó, D.I. Svergun, R. Tauler, M. Pons, Low-resolution structures of transient protein-protein complexes using small-angle X-ray scattering, Journal of the American Chemical Society. 131 (2009) 4378–4386.
[47] T.E. Williamson, B.A. Craig, E. Kondrashkina, C. Bailey-Kellogg, A.M. Friedman, Analysis of self-associating proteins by singular value decomposition of solution scattering data, Biophysical Journal. 94 (2008) 4906–4923.
[48] MATLAB version 7.8.0.347, Natick, Massachusetts: The MathWorks Inc. (2009).
[49] E. Mashiach-Farkash, R. Nussinov, H.J. Wolfson, SymmRef: A flexible refinement method for symmetric multimers, Proteins. (2011).
[50] Y. Liu, G. Gotte, M. Libonati, D. Eisenberg, Structures of the two 3D domain-swapped RNase A trimers., Protein Science : a Publication of the Protein Society. 11 (2002) 371–80.
[51] K. Volz, A test case for structure-based functional assignment: the 1.2 A crystal structure of the yjgF gene product from Escherichia coli., Protein Science : a Publication of the Protein Society. 8 (1999) 2428–37.
[52] B.W. Segelke, D. Nguyen, R. Chee, N.H. Xuong, E.A. Dennis, Structures of two novel crystal forms of Naja naja naja phospholipase A2 lacking Ca2+ reveal trimeric packing., Journal of Molecular Biology. 279 (1998) 223–32.
[53] S. Bressanelli, K. Stiasny, S.L. Allison, E.A. Stura, S. Duquerroy, J. Lescar, et al., Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation., The EMBO Journal. 23 (2004) 728–38.
[54] R. Loris, D. Tielker, K.-E. Jaeger, L. Wyns, Structural basis of carbohydrate recognition by the lectin LecB from Pseudomonas aeruginosa., Journal of Molecular Biology. 331 (2003) 861–70.
[55] J. Badger, J.M. Sauder, J.M. Adams, S. Antonysamy, K. Bain, M.G. Bergseid, et al., Structural analysis of a set of proteins resulting from a bacterial genomics project., Proteins. 60 (2005) 787–96.

[56] C. Hamiaux, J. Pérez, T. Prangé, S. Veesler, M. Riès-Kautt, P. Vachette, The BPTI decamer observed in acidic pH crystal forms pre-exists as a stable species in solution., Journal of Molecular Biology. 297 (2000) 697–712.
[57] T. Stehle, S.C. Harrison, High-resolution structure of a polyomavirus VP1-oligosaccharide complex: implications for assembly and receptor binding., The EMBO Journal. 16 (1997) 5139–48.
[58] R.I. Dima, D. Thirumalai, Asymmetry in the shapes of folded and denatured states of proteins, The Journal of Physical Chemistry B. 108 (2004) 6564–6570.
[59] N. Rawat, P. Biswas, Shape, flexibility and packing of proteins and nucleic acids in complexes, Physical Chemistry Chemical Physics. 13 (2011) 9632–9643.
[60] R. Méndez, R. Leplae, L. De Maria, S.J. Wodak, Assessment of blind predictions of protein-protein interactions: current status of docking methods., Proteins. 52 (2003) 51–67.
[61] S.J. de Vries, A.M.J.J. Bonvin, CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK, PLoS ONE. 6 (2011) e17695.
[62] A. Melquiond, E. Karaca, P.L. Kastritis, A. Bonvin, Next challenges in protein-protein docking: from proteome to interactome and beyond, Wiley Interdisciplinary Reviews: Computational Molecular Science. 2 (2011) 642–651.
[63] C.V. Robinson, A. Sali, W. Baumeister, The molecular sociology of the cell., Nature. 450 (2007) 973–82.
[64] A.T. Brunger, P.D. Adams, L.M. Rice, Annealing in crystallography: a powerful optimization tool., Progress in Biophysics and Molecular Biology. 72 (1999) 135–55.
[65] D. Muradov, B. Kobe, E.N. Dixon, T. Huber, Hybrid Methods for Protein Structure Prediction, in: Hybrid Methods for Protein Structure Prediction, John Wiley & Sons, 2010: pp. 265–277.
[66] T.L. Pukala, B.T. Ruotolo, M. Zhou, A. Politis, R. Stefanescu, J.A. Leary, et al., Subunit architecture of multiprotein assemblies determined using restraints from gas-phase measurements, Structure. 17 (2009) 1235–1243.
[67] F. Gabel, A simple procedure to evaluate the efficiency of bio-macromolecular rigid-body refinement by small-angle scattering, European Biophysics Journal. 41 (2012) 1–11.
[68] N. Nishimura, K. Hitomi, A.S. Arvai, R.P. Rambo, C. Hitomi, S.R. Cutler, et al., Structural mechanism of abscisic acid binding and signaling by dimeric PYR1., Science. 326 (2009) 1373–9.
[69] N. Pinotsis, S. Lange, J.-C. Perriard, D.I. Svergun, M. Wilmanns, Molecular basis of the C-terminal tail-to-tail assembly of the sarcomeric filament protein myomesin., The EMBO Journal. 27 (2008) 253–64.
[70] W.D. Schubert, C. Urbanke, T. Ziehm, V. Beier, M.P. Machner, E. Domann, et al., Structure of internalin, a major invasion protein of Listeria monocytogenes, in complex with its human receptor E-cadherin., Cell. 111 (2002) 825–36.
[71] O. Pertz, D. Bozic, A.W. Koch, C. Fauser, A. Brancaccio, J. Engel, A new crystal structure, Ca2+ dependence and mutational analysis reveal molecular details of E-cadherin homoassociation., The EMBO Journal. 18 (1999) 1738–47.

Chapter

5

# Application Examples of Integrative Modeling

Unraveling the structural basis of Josephin selectivity in poly-ubiquitin cleavage and Structural insights into a H3-H4 histone-chaperones exchange complex in nucleosome formation

Partly based on the research article:
**Guiseppe Nicastro, Sokol V. Todi, Ezgi Karaca, Alexandre M.J.J. Bonvin, Henry L. Paulson, Annalisa Pastore,** Understanding the Role of the Josephin Domain in the PolyUb Binding and Cleavage Properties of Ataxin-3, *PLoS ONE*. 5 (2010) e12430.


**Abstract**

Integrative approaches that incorporate various biophysical and/or biochemical information into the modeling process can help exploring biological function at atomic detail or suggest hypothesis-driven experiments. They are particularly valuable when conventional high-resolution structural determination methods fail. Within this context, this chapter presents two examples of integrative modeling using our information-driven docking approach HADDOCK. In the first, using NMR chemical shift perturbation and mutagenesis data, we show that the topology of the deubiquitination enzyme Josephin imposes a preference for the cleavage of K48-linked di-ubiquitin chains. In the second example, we model, based on the EPR and SAXS data, the transfer complex of the histone heterodimer H3-H4 from Asf1 to p48, two histone chaperones, known to be involved in the initiation of nucleosome assembly formation. The generated models provide insight into the interactions and suggest mutations for further experimental validation.

## 1. Introduction

Numerous biophysical and/or biochemical experimental methods can provide in a reasonably fast manner reliable low-resolution information on biomolecular complexes. This information can become very valuable when the conventional high-resolution structure determination techniques fail. In order to translate these low-resolution data into structural information, efficient integrative modeling tools are needed. These are helpful to address biological function at atomic detail or to reduce the number of possible experiments by suggesting hypothesis-driven ones (see **Chapter 1**). In this chapter, we illustrate integrative modeling with two application examples underpinning the mentioned advantages. In the first, using NMR and mutagenesis data, we investigate the structural basis of a deubiquitination enzyme's biological activity. In the second, we model, based on EPR and SAXS data, the complex between histones H3-H4 and two histone-chaperones leading to initiation of the nucleosome assembly. For this, we use our information-driven docking program HADDOCK [1,2]. In the first example, we follow a *Flexible Multidomain Docking* (FMD) protocol (see **Chapter 3**). In the second case, we make use of the combined HADDOCK$_{SAXS}$ scoring presented in **Chapter 4**.

## 2. The Role of the Josephin Domain in the poly-Ubiquitin Binding and Cleavage properties of Ataxin-3



Ubiquitin (Ub) is an adaptor protein functioning in various pathways, such as transcription, protein degradation, immune response etc. It is attached to protein substrates either as a single Ub molecule or as poly-ubiquitin (polyUb) chains [3]. The latter are formed by linkage of a lysine residue in Ub to the C-terminal of another Ub molecule. Ubiquitin carries seven lysine residues, allowing thus for various linkages. Each linkage type has a different function, among which K48- and K63-linkages are the most studied ones [4]. K48-linked polyUb chains function in the degradation pathway [5–7], whereas K63-linked polyUb chains have a non-degradative signaling role in transcription activation, DNA repair and damage response [8,9]. Detailed knowledge on how differently branched polyUb chains are recognized and differentiated remains elusive [10]. So far, it has been proposed that the recognition specificity is modulated by proteins consisting of multiple domains with different Ub interaction motifs. Ataxin-3, which takes part in quality control pathways and transcriptional regulation, is such a protein: it preferentially cleaves K48-, K63- and mixed polyUb chains (with four or more subunits) [11–14]. Ataxin-3 entails a catalytic N-terminal Jospehin domain that has a cysteine protease fold (**Figure 1A**). Chemical Shift Perturbation (CSP) NMR experiments has unveiled two

distinct Ub binding sites on Josephin: a proximal (close to the active site, site 1) and a distal one (site 2) (**Figure 1A**) [15]. The structure of Josephin bound to polyUb chains is however unknown. In this work, we explore by information-driven docking with HADDOCK, under the guidance of mutation and CSP-NMR data, whether the topology of Josephin imposes any preference for differently branched (K48- or K-63) diUb linkages [10].

### *2.1. Data used to drive the integrative modeling of the Josephin-diUb complex*

The starting structures for Josephin and Ub were taken from PDB entries 1yzb (NMR ensemble) [16] and 1ubq (X-Ray structure) [17], respectively. Ambiguous Interaction Restraints (AIRs) were defined from CSP data obtained from NMR titrations of Josephin constructs with a mutation at either site 1 (I77K-Q78K) or site 2 (W87K). SAMPLEX [18] was run to determine the significant changes in the CSP data of the two binding sites. The *active* residues for docking were defined based on the SAMPLEX selection. For Ub, the *passive* residues were defined as the solvent accessible neighbors of the *active* ones within a 5Å cut-off, whereas for Josephin this cut-off was set to 10Å, in order to increase the interface coverage on Josephin [10].

### *2.2. Description of the integrative modeling strategy*

Three different three-body (two Ub molecules and Josephin) docking runs were performed to test the ability of K48- and K63-linked diUb to interact with Josephin [10]. The isopeptide bond between the two Ubs was defined by imposing a loose distance restraint of 10 Å between the side chain ammonium group of either K48 or K63 and the carboxyl carbon atom of G76 for the initial docking; this distance was subsequently decreased to 1.5Å for the final water refinement to reestablish the isopeptide bond (as described in **Chapter 3**). In the first two runs, the K48- and K63-linkages were imposed separately, whereas in the third run an ambiguous linkage restraint was imposed to both K48 and K63 residues to see which one would be preferentially selected during docking. In the third docking run, an ambiguous distance was defined to both K48 and K63. The C-terminal tail (last three residues) of each Ub was defined as fully flexible together with K48 and/or K63 depending on the linkage type to allow for increased conformational sampling both across the linkage and in the C-terminal tail that should get into proximity of the active site. In all three runs, 5000 models were generated by rigid body docking (*it0*) and the best 200 were subjected to a semi-flexible refinement in torsion angle space (*it1*) followed by refinement in explicit water (*water*). Random removal of restraints was turned off (noecv=false) and additional *center-of-mass* restraints were defined between the

various molecules to ensure the compactness of the solutions. All other parameters were left to their HADDOCK default values.

All three runs were repeated with one additional distance restraint (10 Å) between the Josephin's active site (side chain sulphur of Cys14) and the Ub's C-terminal carbon not involved in the diUb linkage (**Figure 1**). This distance was shortened to 5Å for the final *water* refinement step. The final models were clustered based on the pairwise ligand interface RMSD with a minimum cluster size of four and a RMSD cut-off of 5Å (this value was decreased from the default 7.5Å to have a better clustering in this multibody docking case). The resulting clusters were ranked based on the average HADDOCK score of their top four members.

### *2.3. The topology of Josephin selects preferentially for K48-linked diUb linkages*

The docking run with the K48-diUb linkage in combination with the ambiguous interaction restraints (AIRs) defined from the CSP data resulted in two ensembles of solutions with similar scores (data not shown). Site 1 (or proximal) Ub shares the same orientation in all solutions, suggesting that this site is overall better defined. For Ub binding to site 2 (distal), two contiguous binding surfaces were found that seem to be equally compatible with the experimental restraints [10]. Strikingly, both clusters contain solutions with the C- terminus of site 1 Ub at close proximity to the Josephin active site, even though no explicit distance restraints were defined to position it close to the active site of Josephin (**Figure 1A**). In the second run with the K63 Ub linkage, the solutions were more scattered and no biologically meaningful structures could be obtained (i.e. none having the Ub C-terminus close to the catalytic center) [10]. Even in the best case, the C-terminus of site 1 Ub was far away from the active site (**Figure 1B**). To investigate if this could be an artifact of our docking procedure, we repeated the calculations with an additional distance restraint to position the C-terminus of the site 1 Ub into the Josephin active site. No solution was found that could satisfy this restraint indicating that the two subunits of a K63-linked diUb cannot be accommodated simultaneously in both Josephin sites. Finally, we performed a run in which the linkage preference was left ambiguous by defining a linkage restraint including both K48 and K63. Although both linkage types were obtained at the rigid-body docking stage, the only selected solutions for the subsequent semi-flexible refinement based on the HADDOCK score correspond to the K48-linkage. These results indicate that the different linkages result in different positioning of the two Ubs on Josephin and dictate the different binding specificities: The K63-linked diUb can not accommodate the geometric requirements imposed by the two Josephin binding sites and only the K48-linkage diUb is able to occupy both Ub-binding sites of Josephin simultaneously. This observation was further supported

by biochemical experiments. Our collaborators incubated isolated Josephin with equal amounts of K48-linked or K63-linked pentaUb chains. The isolated Josephin domain cleaved K48-linked pentaUb chains with a higher efficiency compared to K63-linked chains, which indicates a preference of Josephin for K48-linkages (**Figure 1C**) [10], consistent with our prediction based on the docking models.

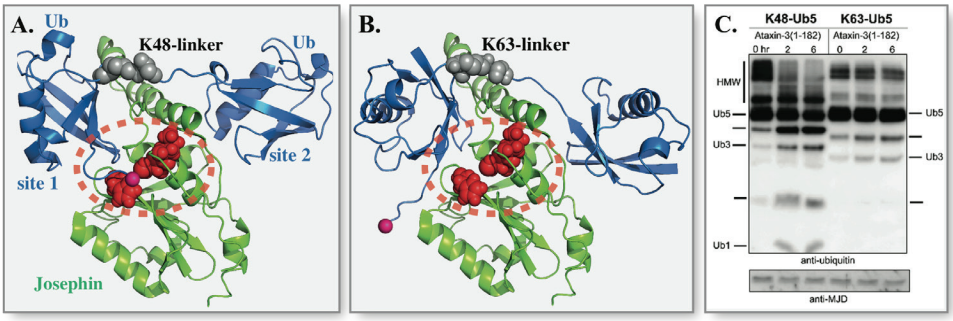



**Figure 1. NMR CSP data-driven HADDOCKing unveils that Josephin's topology dictates a preference for K48-linked diUb chains, This is further supported by enzymatic assays.** Josephin is depicted in green cartoon and its encircled active site (C14, H119, N134) by red spheres. diUb is represented as marine-blue cartoon, its C-terminus as magenta sphere and its linked lysines as green spheres. The two subunits of K48-linked diUb could populate the binding and catalytic sites on Josephin simultaneously (**A**), whereas K63-linked diUb chains could not (**B**). This suggests a preference for the cleavage of K48-linked poly-ubiquitin chains, which was demonstrated experimentally (**C**): Equal amounts of pentaUb K48- or K63-linked chains were incubated together with Josephin, which revealed that isolated Josephin domain cleaves K48-linked pentaUb chains with a higher efficiency compared to K63-linked ones.



## 3. Structural Basis for the H3-H4 histone Exchange Complex between two Different Histone-chaperons, p48 and Asf1

The nucleosome is the integral subunit of the eukaryotic chromatin structure. Its core is composed of an octomeric histone complex that is formed by association of, first, two H3-H4 heterodimers with DNA and, then, two H2A-H3B heterodimers with the formed histone-DNA complex [19,20]. During DNA replication, histone chaperons assist the formation of nucleosomes and recruits enzymes to them for post-translational modification [21]. Retinoblastoma protein RbAp48 (p48) is one of the key chaperones in this process. Its structure belongs to the WD40 repeat β-propeller fold (**Figure 2B**) [22,23] and is a subunit of the chromatin-assembly-factor-1 complex (CAF-1) that positions heterodimeric H3-H4 histones onto newly replicated DNA to initiate nucleosome assembly [21,24,25]. During the initiation process, CAF-1 is assisted by another critical chaperone, anti-silencing function protein 1 (Asf1) [26].

The crystal structure of H3-H4-Asf1 showed that Asf1 envelops H3 to prevent H3-H4 oligomerization (Asf1, **Figure 2A**), facilitating the transfer of H3-H4 to CAF-1 as a dimer. Here, based on SAXS and EPR data combined with docking, we investigate, the structural basis of the initiation mechanism of nucleosome assembly formation, i.e. the p48-H3-H4-Asf1 complex that captures the transfer of H3-H4 heterodimer from Asf1 to p48.

### *3.1. Data used to drive the integrative modeling of p48-H3-H4-Asf1*

For modeling the p48-H3-H4-Asf1 complex, two starting structures were used. The first one corresponds to a model of the complex between p48 (pdb id: 2xu7 [27]) and the N-terminal helix of H4 histone (p48-H4$^{Nter}$) based on the crystal structure of the p46-H4$^{Nter}$ complex, 3cfs [21] (**Figure 2B**). The key residues involved in hydrophobic or hydrophilic/charged interactions with H4 and defining the protein fold are identical in p48 and p46 (the overall sequence homology is 90%) [21]. The model was constructed as follows: p48 was first rigidly fit on p46-H4$^{Nter}$ complex (in Pymol [28]), then p48 and H4$^{Nter}$(taken from the p46-H4$^{Nter}$ complex) were extracted and the resulting complex was optimized with the refinement server of HADDOCK. The second structure used for docking was taken from the crystal structure of the Asf1-H3-H4 complex (H3-ΔH4-Asf1) [26] with the N-terminal helix of H4 chopped off.

SAXS data (collected on a 1.0 mg/ml sample) (Aleksandra Watson, Cambridge University, unpublished data) and distance information from PELDOR EPR experiments were available for modeling the p48-H3-H4-Asf1 complex. The SAXS profile was used for scoring the docking models (see **Chapter 4**). The EPR data were not used as restraints, but were indicative of structural rearrangements during complex formation (The distance between the labeled N-terminal helices of H3 and H4 was observed to dramatically change upon H3-H4's complexation with p48, Ernest Laue, Cambridge University, unpublished data). This information was used to devise a docking protocol allowing for increased flexibility (see below).

### *3.2. Description of the integrative modeling strategy*

In order to explore the structural grounds of the conformational change suggested by EPR experiments, the H3-H4 heterodimer (extracted from 2hue) was fit onto the H4$^{Nter}$ of the p48-H4$^{Nter}$ model. This rigid-body fitting resulted in serious steric clashes (**Figure 2C**), indicating that H4 should experience major conformational changes when it binds to p48. In order to generate possible models of the p48-H3-H4-Asf1 complex allowing for conformational changes, we docked p48-H4$^{Nter}$ and H3-ΔH4-Asf1, using only center-of-mass restraints in combination with a

loose unambiguous restraint of 10 Å between G41 and G42 where the cut in H3-H4 was introduced. This distance was reduced to 1.3 Å for the final *water* refinement. In addition, three residues from each side of the separated H4 structural segments ($H4^{Nter}$ and ΔH4) were treated as fully flexible. Their terminus was kept uncharged. The number of structures at the *it0*, *it1* and *water* stages was increased to 10,000/400/400 respectively. At the end of *it0*, models were selected for further refinement by using the combined $HADDOCK_{SAXS}$ score (see Chapter 4). All other parameters were left to their HADDOCK default values. The final models were clustered with a clustering cut-off of 7.5Å and ranked based on the average HADDOCK score of their top four members.

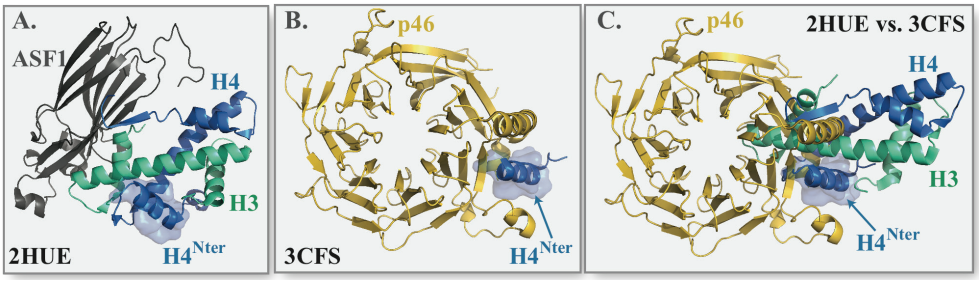



**Figure 2. The superimposed N-terminal helices of histone H4 belonging to 2hue (H3-H4-Asf1) and 3cfs (p46-H3-H4) suggest that a dramatic structural rearrangement must occur in H4 upon complexation with p48 (a close homologue of p46).** Asf1 is illustrated in black cartoon, H3 in cyan-green cartoon, H4 in marine-blue cartoon and p46 in gold cartoon. The N-terminal helix of H4 ($H4^{Nter}$) is represented with marine-blue transparent surface. **A.** Asf1 envelopes H3, in order to block H3-H4 heterotetramer formation [26]. **B.** p46 contains a groove formed between N-terminal helix and an extended loop, to which $H4^{Nter}$ binds [21]. **C.** H4 cannot preserve its conformational state captured in 2hue, as superimposition of it on top of 3cfs's $H4^{Nter}$ results in serious steric clashes.

### *3.3. Initial models of p48-H3-H4-Asf1 exchange complex suggest further mutational studies*

The docking procedure described in **Section 3.2** resulted in three clusters (**Table 1**, **Figure 3**). Although the binding mode of each cluster is different they all (i) have p48 and Asf1 in close vicinity to facilitate the transfer of H3-H4 heterodimer, (ii) fit the experimental SAXS data equally well with comparable χ values and (iii) in all, the termini of $H4^{Nter}$ and ΔH4 are close enough to be re-connected. Their HADDOCK scores, however, differ significantly with Cluster 1 having the best score (and best individual components as well). We therefore selected it as the most plausible binding mode (**Figure 3A**).

**Table 1. The statistics of individual scores on a per-cluster basis[a].**

| | HS (a.u.) | SAXS-χ | $E_{vdW}$ (kcal/mol) | $E_{elec}$ (kcal/mol) | $E_{desol}$ (kcal/mol) | BSA (Å$^2$) |
|---|---|---|---|---|---|---|
| **Cluster 1** | -59.1±2.5 | 1.03±0.01 | -32.4±4.2 | -250±29 | 16.0±8.9 | 1447±84 |
| **Cluster 2** | -19.7±5.9 | 1.04±0.00 | -22.9±1.4 | -182±13 | 32.5±7.8 | 1090±48 |
| **Cluster 3** | -4.0±14.0 | 1.01±0.00 | -29.8±1.0 | -49±28 | 25.6±13.0 | 1122±84 |

HS corresponds to HADDOCK Score, $E_{vdw}$ to van der Waals energy, $E_{elec}$ to Electrostatics energy, $E_{desol}$ to Desolvation energy and BSA to Buried Surface Area, [a]The clusters were ranked based on the average HADDOCK score (1.0 $E_{vdw}$ + 0.2 $E_{elec}$ + 1.0 $E_{desol}$) of the top four members of a cluster.

We further analyzed Cluster 1 to suggest critical residues across the p48-H4$^{Nter}$/H3-ΔH4-Asf1 interface that can be targeted for further mutagenesis studies (**Table 2**). The interaction surface on H3-ΔH4-Asf1 is majorly formed by the middle-long helix of H3, with a contribution by a small linker region of Asf1 facing towards this H3 helix (**Figure 3A**). On the p48 side, the interaction surface is mainly formed by the N-terminal helix of p48. Overall, the interface residues are mainly composed of highly charged and hydrophobic residues that lead to a low electrostatics energy and a large buried surface area. We analyzed the interface of all three clusters (**Table 2A**) and, based on this, we could propose amino-acids for mutagenesis that are involved in contacts unique to Cluster 1 (**Table 2A**). These can be used as candidates to validate the model by further mutagenesis studies (**Table 2B**). We should note here that, at the time of writing, this is still an ongoing project and hopefully more data will become available to increase the quality and the precision of the models generated.

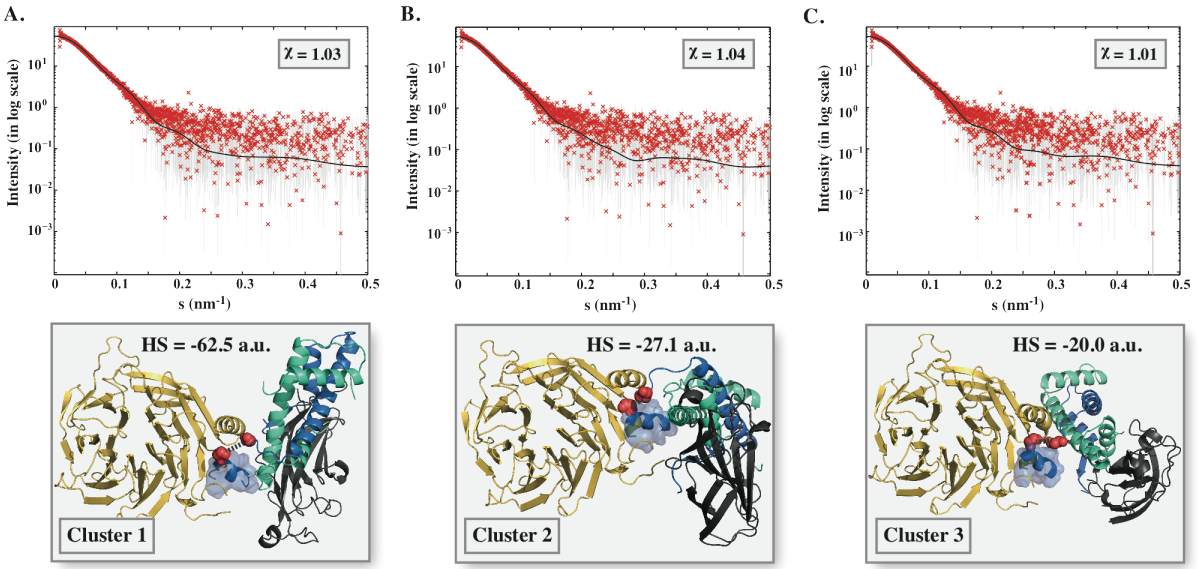



**Figure 3. The three p48-H3-H4-Asf1 docking models show different plausible binding modes capturing the suggested conformational change of H4 histone.** Cluster 1 (**A**) has the best HADDOCK score (see **Table 1**). All models have p48 and Asf1 in close vicinity to facilitate the transfer of H3-H4 heterodimer. Furthermore, the termini of $H4^{Nter}$ and ΔH4 (the separated components of H4) are close enough in space to be re-connected (depicted in red spheres). All three models fit well the experimental SAXS profile with χ's varying between 1.01 and 1.04. The fit was performed using Crysol [29].3-H4-Asf1 docking models show different plausible binding modes capturing the suggested conformational change of H4 histone. Cluster 1 (**A**) has the best HADDOCK score (see **Table 1**). All models have p48 and Asf1 in close vicinity to facilitate the transfer of H3-H4 heterodimer. Furthermore, the termini of $H4^{Nter}$ and ΔH4 (the separated components of H4) are close enough in space to be re-connected (depicted in red spheres). All three models fit well the experimental SAXS profile with χ's varying between 1.01 and 1.04. The fit was performed using Crysol [29].

**Table 2A. List of the contacts observed across the p48-H4$^{Nter}$/H3-ΔH4-Asf1 interface for all three clusters[a].**

| p48-H3 Interface | Hydrogen Bonds[b] | | | Non-bonded contacts | Cluster id |
|---|---|---|---|---|---|
| | M-M | S-S | M-S | | |
| LYS26 - GLU105 | | x | | x | 1 |
| LYS26 - VAL101 | | | | x | 1 |
| PRO29 - LEU61 | | | | x | 3 |
| PHE30 - LEU61 | | | | x | 2 |
| PHE30 - PHE104 | | | | x | 3 |
| PHE30 - VAL101 | | | | x | 3 |
| VAL35 - LEU61 | | | | x | 3 |
| **p48-H4 interface** | | | | | |
| GLU19-LYS44 | | x | | x | 2 |
| LYS26-ILE46 | | | | x | 3 |
| LYS26-LYS44 | | | | x | 3 |
| LYS26-VAL43 | | | | x | 3 |
| ASN27-GLY42 | | | x | x | 3 |
| PHE30-TYR98 | | | | x | 1 |
| **p48-Asf1 interface** | | | | | |
| PHE30-LYS143 | | | x | x | 1 |
| **H4$^{Nter}$-Asf1 interface** | | | | | |
| LYS31-GLU142 | | x | x | x | 1 |
| ILE34-GLU142 | | | | x | 1 |
| **H4$^{Nter}$-H3 interface** | | | | | |
| LYS31-GLU94 | | x | | | 2 |
| PRO32-ALA98 | | | | x | 2 |
| ALA33-GLU97 | | | | x | 2 |
| ALA33-ALA98 | | | | x | 2 |
| ALA33-VAL101 | | | | x | 2 |
| ALA33-ILE112 | | | | x | 1 |
| ARG36-GLU105 | | x | | x | 2 |
| ARG36-ILE112 | | | x | | 1 |
| ARG36-LYS115 | | | x | x | 1 |
| LEU37-ILE112 | | | | x | 1 |
| LEU37-LEU61 | | | | x | 2 |
| ARG40-GLU105 | | x | | | 2 |

**Table 2A continued:** [a]The intermolecular contacts were calculated on the best four members of each cluster. Only contacts observed in at least in 2 models among the best four member of each cluster are listed. The known contacts across the p48-H4$^{Nter}$ interface were excluded from the list, [b]M-M is indicates Main chain-Main chain, S-S Side chain-Side chain, M-S Main chain-Side chain.

**Table 2B. Possible critical residues (across the p48-H4$^{Nter}$/H3-ΔH4-Asf1 interface) suggested for further mutagenesis studies[c].**

| p48 | H3 | H4 | Asf1 |
|---|---|---|---|
| GLU19<br>LYS26<br>ASN27<br>PHE30 | GLU105<br>ILE112 | LYS31<br>ALA33<br>ARG36<br>ARG40<br>LYS44 | LYS142<br>GLU143 |

[c]The selection criteria for choosing a critical residue pair were, (i) to be involved in a contact unique to Cluster 1, (ii) to be observed in at least in 2 models among the best four member of Cluster 1 and (iii) to significantly contribute to the intermolecular energy ($E_{vdw}$+$E_{elec}$).

## 4. Concluding Remarks

In this chapter, we have demonstrated the use of integrative modeling with two challenging examples, applying the protocols and tools reported in the previous chapters of this thesis. For this, low-resolution data were transformed into structural information, in order to explore the grounds of the observed biological activity. The models obtained in both cases should not be seen as an end, but rather as a starting point for new hypothesis that can be tested experimentally. Within this context, in the first example, the modelling allowed to predict the cleavage preference of Josephin for differently branched diUb chains, which was subsequently validated by biochemical experiments, and in the second one, new hypothesis-driven mutagenesis experiments were proposed to validate the structural model of H3-H4's transfer from Asf1 to p48. As long as high-resolution structural studies of this complex fail, further validation, for example by targeted mutagenesis of the proposed models will be required, as the SAXS data alone are not sufficient to distinguish between various binding modes. This work also nicely illustrates the fact that a tight collaboration between computational structural biologist and experimentalists is the mainstay of integrative modeling. Joining forces can lead to an understanding of complex biological phenomena that cannot be thoroughly grasped either by modeling or experiment only.

## References

[1] S.J. de Vries, M. van Dijk, A.M.J.J. Bonvin, The HADDOCK web server for data-driven biomolecular docking, Nature Protocols. 5 (2010) 883–897.
[2] C. Dominguez, R. Boelens, A.M.J.J. Bonvin, HADDOCK: a protein-protein docking approach based on biochemical or biophysical information., Journal of the American Chemical Society. 125 (2003) 1731–7.
[3] V. Chau, J.W. Tobias, A. Bachmair, D. Marriott, D.J. Ecker, D.K. Gonda, et al., A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein., Science. 243 (1989) 1576–83.
[4] A. Hershko, A. Ciechanover, The ubiquitin system., Annual Review of Biochemistry. 67 (1998) 425–79.
[5] E.S. Johnson, P.C. Ma, I.M. Ota, A. Varshavsky, A proteolytic pathway that recognizes ubiquitin as a degradation signal., The Journal of Biological Chemistry. 270 (1995) 17442–56.
[6] C.M. Pickart, Targeting of substrates to the 26S proteasome., Journal of Federation of American Societies for Experimental Biology. 11 (1997) 1055–66.
[7] A. Varshavsky, Regulated protein degradation., Trends in Biochemical Sciences. 30 (2005) 283–6.
[8] C.M. Pickart, D. Fushman, Polyubiquitin chains: polymeric protein signals., Current Opinion in Chemical Biology. 8 (2004) 610–6.
[9] A. Ciechanover, The ubiquitin-proteasome proteolytic pathway., Cell. 79 (1994) 13–21.
[10] G. Nicastro, S.V. Todi, E. Karaca, A.M.J.J. Bonvin, H.L. Paulson, A. Pastore, Understanding the Role of the Josephin Domain in the PolyUb Binding and Cleavage Properties of Ataxin-3, PLoS ONE. 5 (2010) e12430.
[11] O. Riess, U. Rüb, A. Pastore, P. Bauer, L. Schöls, SCA3: neurological features, pathogenesis and animal models., Cerebellum. 7 (2008) 125–37.
[12] B. Burnett, F. Li, R.N. Pittman, The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity., Human Molecular Genetics. 12 (2003) 3195–205.
[13] Y. Chai, S.L. Koppenhafer, S.J. Shoesmith, M.K. Perez, H.L. Paulson, Evidence for proteasome involvement in polyglutamine disease: localization to nuclear inclusions in SCA3/MJD and suppression of polyglutamine aggregation in vitro., Human Molecular Genetics. 8 (1999) 673–82.
[14] B.J. Winborn, S.M. Travis, S.V. Todi, K.M. Scaglione, P. Xu, A.J. Williams, et al., The deubiquitinating enzyme ataxin-3, a polyglutamine disease protein, edits Lys63 linkages in mixed linkage ubiquitin chains., The Journal of Biological Chemistry. 283 (2008) 26436–43.
[15] G. Nicastro, L. Masino, V. Esposito, R.P. Menon, A. De Simone, F. Fraternali, et al., Josephin domain of ataxin-3 contains two distinct ubiquitin-binding sites., Biopolymers. 91 (2009) 1203–14.
[16] G. Nicastro, R.P. Menon, L. Masino, P.P. Knowles, N.Q. McDonald, A. Pastore, The solution structure of the Josephin domain of ataxin-3: structural determinants for molecular recognition., Proceedings of the National Academy of Sciences of the United States of America. 102 (2005) 10493–8.
[17] S. Vijay-Kumar, C.E. Bugg, W.J. Cook, Structure of ubiquitin refined at 1.8Å resolution., Journal of Molecular Biology. 194 (1987) 531–44.
[18] M. Krzeminski, K. Loth, R. Boelens, A.M.J.J. Bonvin, SAMPLEX: automatic mapping of perturbed and unperturbed regions of proteins and complexes., BMC Bioinformatics. 11 (2010) 51.
[19] C. Gruss, J. Wu, T. Koller, J.M. Sogo, Disruption of the nucleosomes at the replication fork., The EMBO Journal. 12 (1993) 4533–45.

[20] K. Luger, A.W. Mäder, R.K. Richmond, D.F. Sargent, T.J. Richmond, Crystal structure of the nucleosome core particle at 2.8 A resolution., Nature. 389 (1997) 251–60.
[21] N.V. Murzina, X.-Y. Pei, W. Zhang, M. Sparkes, J. Vicente-Garcia, J.V. Pratap, et al., Structural basis for the recognition of histone H4 by the histone-chaperone RbAp46., Structure. 16 (2008) 1077–85.
[22] M.R. Parthun, J. Widom, D.E. Gottschling, The major cytoplasmic histone acetyltransferase in yeast: links to chromatin replication and histone metabolism., Cell. 87 (1996) 85–94.
[23] A. Verreault, P.D. Kaufman, R. Kobayashi, B. Stillman, Nucleosomal DNA regulates the core-histone-binding subunit of the human Hat1 acetyltransferase., Current Biology : CB. 8 (1998) 96–108.
[24] A. Verreault, P.D. Kaufman, R. Kobayashi, B. Stillman, Nucleosome assembly by a complex of CAF-1 and acetylated histones H3/H4., Cell. 87 (1996) 95–104.
[25] S. Smith, B. Stillman, Purification and characterization of CAF-I, a human cell factor required for chromatin assembly during DNA replication in vitro., Cell. 58 (1989) 15–25.
[26] C.M. English, M.W. Adkins, J.J. Carson, M.E.A. Churchill, J.K. Tyler, Structural basis for the histone chaperone activity of Asf1., Cell. 127 (2006) 495–508.
[27] S. Lejon, S.Y. Thong, A. Murthy, S. AlQarni, N.V. Murzina, G.A. Blobel, et al., Insights into association of the NuRD complex with FOG-1 from the crystal structure of an RbAp48·FOG-1 complex., The Journal of Biological Chemistry. 286 (2011) 1196–203.
[28] The PyMOL Molecular Graphics System, Schrödinger, LLC. (n.d.).
[29] D. Svergun, C. Barberato, M.H.J. Koch, CRYSOL-a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates, Journal of Applied Crystallography. 28 (1995) 768–773.

Chapter

6

# Perspectives: On the Future and Limitations of Integrative Modeling

*"We have entered the era of an integrated structural biology"*
(From the meeting review of the symposium celebrating the 40th anniversary of the Protein Data Bank archive)

Based on two review articles:
**Adrien S.J. Melquiond, Ezgi Karaca, Panagiotis L. Kastritis, Alexandre M.J.J. Bonvin**, Next challenges in protein–protein docking: from proteome to interactome and beyond. *WIREs Comput Mol Sci* 2012, 2: 642-651.
**Ezgi Karaca, Alexandre M.J.J. Bonvin**, Advances in Integrative Modeling of Biomolecular Complexes, *Methods*, 2012, *advances online publication*.

A mechanistic understanding of how a cell functions involves complementing interactomes and cellular tomograms by three-dimensional structures of complexes. This daunting task can only be achieved by joining experimentation and computational modeling. Integrative modeling techniques that can judiciously combine and accurately translate sparse experimental data into structural information are excellent examples, which have been developed to this end (**Figure 1**). In the last decade, following the advances in biophysics and biochemistry, the coverage of the integrative approaches has increased substantially, such that acquiring structural information on challenging macromolecular complexes, like the proteasome, the ribosome or the Nuclear Pore Complex has come within the reach of integrative modeling.

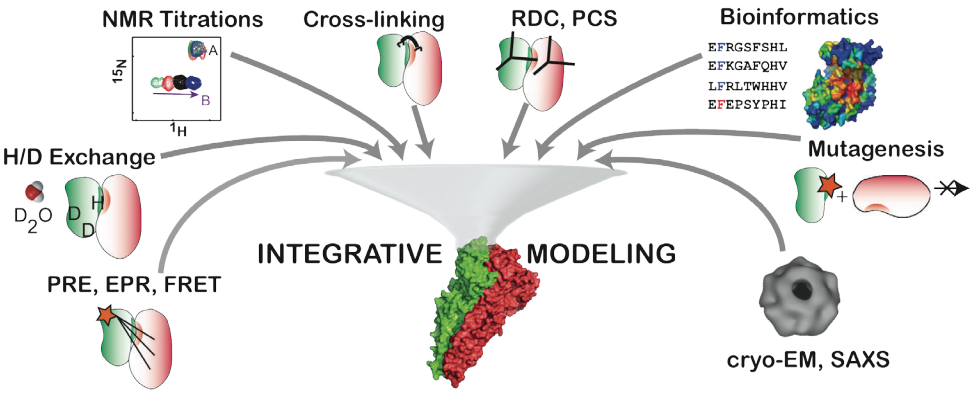



**Figure 1. Integrative modeling can translate low-resolution and sparse *biophysical, biochemical* and *bioinformatics* data into structural information.** The supplied information can be handled either *a priori* during sampling, by restraining the conformational space, or *a posteriori* during scoring, by filtering or ranking the generated models based on their fit to the experimental data.

Most of the current macromolecular docking programs, with HADDOCK as one of the pioneers, have stepped into the integrative modeling field. The recent and ongoing developments in docking, together with the ones presented in this thesis, are important milestones in the prediction of challenging macromolecular complexes. However, further methodological advancements are required to move to the next level of complexity, i.e. building accurate and comprehensive models of larger machineries, such as, for example, the bacterial flagellar motor, the atp synthase, the translocon or molecular transporters.

A possible avenue for this is to incorporate low-resolution shape data, coming from cryo-EM (see **Chapter 1**) or SAXS, not only in the scoring stage to discriminate the generated models, but also in the sampling stage, to drive the refinement process. This will require the definition of energy functions and their derivatives that measure the agreement between the experimental data, SAXS profiles or cryo-EM densities, and the model [1,2]. These can be used in optimization algorithms based, for example, on Monte Carlo approaches that only require an energy function or gradient-based methods such as restrained energy minimization and molecular dynamics [2,3]. In the case of SAXS data, direct refinement against the SAXS curve comes at a price: all-atom gradient-based minimization of large complexes are computationally expensive [1,4]. In order to decrease the computational cost, a simplified version of the energy function can be used during energy minimization [5] or lower resolution information extracted from SAXS profiles, such as the radius of

gyration (Rg) [6] and the maximum molecular dimension, can be used to drive modeling. In this thesis, we have only reported the use of SAXS profiles as a filter for selecting models in combination with the HADDOCK score (**Chapter 4**). A next step in incorporating SAXS data into docking in a computational cost efficient manner is to restrain the radius of gyration of the complex based on Rg obtained from the SAXS data (e.g. from Guinier or Debye approximations or from the integral of the pair distribution function [7]). First tests of Rg-driven docking on the Docking Benchmark 4.0 (results not shown) indicate that such a restraint does not improve the docking performance compared to *ab initio* docking. Therefore, more complex energy functions representing the SAXS data will have to be implemented in order to make a difference.

Accurate modeling of macromolecular machines involves dealing with a large number of molecules, simultaneously, and also extensive sampling of their interaction space, which will lead to a myriad of degrees of freedom (see **Chapter 1**). A possible way to lower the number of degrees of freedom is to decrease the granularity of the system by Coarse Graining (CG) the biomolecular representation. CG can be employed at different resolutions and be incorporated into different stages of modeling, i.e. before or during sampling [8]. An example of incorporating CG into the pre-modeling stage would be to explore the conformational space of the input biomolecule with an Elastic Network Model (ENM) and then perform the docking from an ensemble of structures generated with such methods [9]. On the other hand, during modeling, the entire system, or sub-parts, such as side-chains, residues or even domains, can be represented as beads, in order to employ CG at different resolutions [8,10]. CG is commonly used in the Molecular Dynamics field to simulate mesoscopic systems for longer time scales [11,12]. Moreover, it is used in various molecular docking programs to speed up the sampling process [13,14] or other integrative molecular modeling programs to build up gigantic molecular assemblies, such as the Nuclear Pore Complex [15]. A successful CG implementation in sampling necessitates careful treatment and re-parameterization of the force field according to the coverage of the CG model selected [10]. Therefore, the level of CG, its transferability and the efficiency of the sampling method will directly affect the accuracy level of the resulting model [8,10,16].

So far, most of the docking methods have been developed to model interactions of soluble proteins and consequently overlooked membrane and membrane associated complexes. Membrane proteins constitute an integral part, ~30% of the proteomes of organisms [17], thus it will be crucial to also build robust protocols for accurate modeling of membrane and membrane-associated protein complexes. Here, one issue will be to properly mimic the membrane environment, in order to describe the thermodynamics of the membrane protein complex. This point can be addressed

by refining the model in a medium that simulates the physicochemical properties of the membrane or by incorporating a term into the energy function accounting for the membrane contribution, e.g. a membrane solvation term [18]. Another issue in modeling membrane complexes will be to obtain restraints. These can be provided by solid-state NMR [19], other experimental techniques [7,20–22] or bioinformatics analysis [23]. A recent and promising development is to use evolutionary methods for deriving contact information between residues. This approach has recently been demonstrated to be rather successful in predicting the 3D structure of membrane proteins [24] and has the potential to be extended to complexes of membrane proteins.

Next to computational limitations, the quality of the experimental data used in models. For example, an information-driven method can only produce biologically meaningful models, if the supplied data are guiding the search towards the relevant part of the conformational space. Also, the information provided by the various data should be non-degenerate, i.e. describe distinct properties of the system (see **Chapter 1**). We illustrate this point using cross-linking data derived from MS experiments. Throughout the last decade, method developments and proof-of-concept studies in this area have yielded a wealth of valuable information for the structural analysis of biomolecules, ranging from protein-peptide to multi-protein complexes [25]. We picture here the impact of the degeneracy issue in the modelling of an enzyme-inhibitor complex (Collicin-Immunity protein in complex with collicin dnase, pdb id: 1ujz) for which published cross-linked residue pairs are available [26]: The three inter-monomer distances detected by MS were used to drive the complex formation (**Figure 2A**). The docking resulted in a single cluster corresponding to an alternative binding mode (**Figure 2B** – grey model). The top ranking solution (**Figure 2B** – green model), which did not cluster, was an isolated high accuracy (interface-RMSD < 1 Å) solution. The generation of the alternative binding mode resulted from the fact that two of the supplied inter-monomer distances were sharing the same atom ($Lys^{537}$, **Figure 2A**). To be able to define uniquely the orientation between two surfaces (or two planes) three independent distances are needed. Indeed, when three independent distances evenly distributed over the surface of the complex were used, the docking produced a top-ranking cluster of native solutions with interface-RMSD < 1 Å. This example illustrates that the accuracy of information-driven approaches is directly correlated with the quality and distribution of the input data. It is therefore highly advisable to assess both the degeneracy and the quality of any input data before using them during modeling.

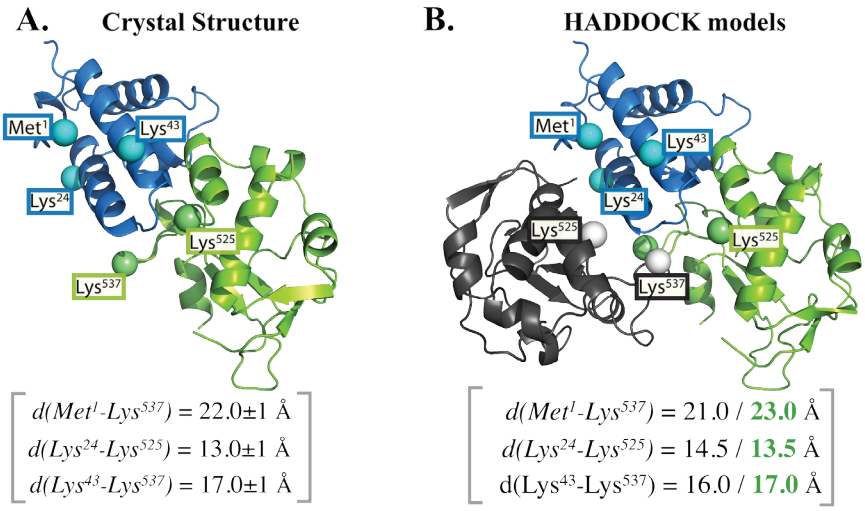



**Figure 2. Dependency of the docking model on the quality and degeneracy of the input data.** Published cross-linked residue pairs for the enzyme-inhibitor complex (Collicin-Immunity protein (marine blue) in complex with collicin dnase (green)) were used to drive the docking process. The cross-linked residues are represented in spheres and labeled with their residue id. The inter-monomer distances used to drive docking are indicated in between brackets. **A.** View of the reference crystal structure (pdb id: 1ujz [27]). **B.** Resulting HADDOCK models: the best ranking cluster, corresponding to an alternative solution, is shown with collicin in black; the isolated (not found in any cluster), top ranking solution, corresponding to the crystal structure, is shown with collicin in green.

In this thesis, we have reviewed the field of integrative modeling and described new approaches that have been implemented in our modeling platform HADDOCK. We anticipate that in the next decade, some of the limitations and challenges laying ahead of us will be addressed, accelerating the growth of the integrative modeling field and paving the way to increase accuracy, coverage and resolution of integrative models. Nevertheless, as in any modeling exercise, we should always keep in mind that the model should not be an endpoint, but rather a starting point for generating new hypothesis that can be tested experimentally. As a final remark, it should be clear that the route to shedding atomic resolution light onto challenging molecular systems and understanding the underlying mechanistic of biomolecular function is not a straight one; its exploration will require computational structural biologists and experimentalists to work in a collaborative and synergistic manner.

## References

[1] T. Madl, F. Gabel, M. Sattler, NMR and small-angle scattering-based structural analysis of protein complexes in solution, Journal of Structural Biology. 173 (2011) 472–482.
[2] F. Alber, F. Förster, D. Korkin, M. Topf, A. Sali, Integrating diverse data for structure determination of macromolecular assemblies., Annual Review of Biochemistry. 77 (2008) 443–477.
[3] A.T. Brunger, P.D. Adams, L.M. Rice, Annealing in crystallography: a powerful optimization tool., Progress in Biophysics and Molecular Biology. 72 (1999) 135–55.
[4] A. Grishaev, J. Wu, J. Trewhella, A. Bax, Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data., Journal of the American Chemical Society. 127 (2005) 16621–8.
[5] F. Gabel, B. Simon, M. Nilges, M. Petoukhov, D. Svergun, M. Sattler, A structure refinement protocol combining NMR residual dipolar couplings and small angle scattering restraints., Journal of Biomolecular NMR. 41 (2008) 199–208.
[6] J. Kuszewski, A.M. Gronenborn, G.M. Clore, Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration, Journal of the American Chemical Society. 121 (1999) 2337–2338.
[7] C.D. Putnam, M. Hammel, G.L. Hura, J.A. Tainer, X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution, Quarterly Reviews of Biophysics. 40 (2007) 191–285.
[8] S.C. Flores, J. Bernauer, S. Shin, R. Zhou, X. Huang, Multiscale modeling of macromolecular biosystems., Briefings in Bioinformatics. 13 (2012) 395–405.
[9] I. Bahar, A.J. Rader, Coarse-grained normal mode analysis in structural biology., Current Opinion in Structural Biology. 15 (2005) 586–92.
[10] V. Tozzini, Coarse-grained models for proteins., Current Opinion in Structural Biology. 15 (2005) 144–50.
[11] A.J. Rader, Coarse-grained models: getting more with less., Current Opinion in Pharmacology. 10 (2010) 753–9.
[12] P. Sherwood, B.R. Brooks, M.S.P. Sansom, Multiscale methods for macromolecular simulations., Current Opinion in Structural Biology. 18 (2008) 630–40.
[13] A. May, M. Zacharias, Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking., Proteins. 70 (2008) 794–809.
Protein Docking, The Journal of Physical Chemistry B. (2011).
[15] D. Russel, K. Lasker, B. Webb, J. Velázquez-Muriel, E. Tjioe, D. Schneidman-Duhovny, et al., Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies., PLoS Biology. 10 (2012) e1001244.
[16] S. Takada, Coarse-grained molecular simulations of large biomolecules., Current Opinion in Structural Biology. 22 (2012) 130–7.
[17] R.M. Bill, P.J.F. Henderson, S. Iwata, E.R.S. Kunji, H. Michel, R. Neutze, et al., Overcoming barriers to membrane protein structure determination., Nature Biotechnology. 29 (2011) 335–40.
[18] A.J. Bordner, B. Zorman, R. Abagyan, Efficient molecular mechanics simulations of the folding, orientation, and assembly of peptides in lipid bilayers using an implicit atomic solvation model., Journal of Computer-aided Molecular Design. 25 (2011) 895–911.
[19] M. Renault, A. Cukkemane, M. Baldus, Solid-state NMR spectroscopy on complex biomolecules., Angewandte Chemie (International Ed. in English). 49 (2010) 8346–57.
[20] R.A. Dilanian, C. Darmanin, J.N. Varghese, S.W. Wilkins, T. Oka, N. Yagi, et al., A new approach for structure analysis of two-dimensional membrane protein crystals using X-ray

powder diffraction data., Protein Science : a Publication of the Protein Society. 20 (2011) 457–64.
[21] N.P. Barrera, C.V. Robinson, Advances in the mass spectrometry of membrane proteins: from individual proteins to intact complexes., Annual Review of Biochemistry. 80 (2011) 247–71.
[22] Werner Kühlbrandt, Combining Cryo-EM and X-ray Crystallography to Study Membrane Protein Structure and Function, Springer Netherlands, Dordrecht, 2012.
[23] T. Nugent, D.T. Jones, Membrane protein structural bioinformatics., Journal of Structural Biology. 179 (2012) 327–37.
[24] T.A. Hopf, L.J. Colwell, R. Sheridan, B. Rost, C. Sander, D.S. Marks, Three-dimensional structures of membrane proteins from genomic sequencing., Cell. 149 (2012) 1607–21.
[25] J. Rappsilber, The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes., Journal of Structural Biology. 173 (2011) 530–40.
[26] J. Seebacher, P. Mallick, N. Zhang, J.S. Eddes, R. Aebersold, M.H. Gelb, Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing., Journal of Proteome Research. 5 (2006) 2270–82.
[27] T. Kortemme, L.A. Joachimiak, A.N. Bullock, A.D. Schuler, B.L. Stoddard, D. Baker, Computational redesign of protein-protein interaction specificity., Nature Structural & Molecular Biology. 11 (2004) 371–9.

# Summary

The process of life, which we see, feel, experience or witness in our daily lives, is essentially modulated at the *nanoscale* ($10^{-9}$ m). Being more precise, the biological function we observe at the *macroscale*, is an outcome of the orchestrated communications among biomolecules -particularly proteins- at the *nanoscale*. Thus, it is by dissecting and grasping the *nanoscale* world of biomolecules and their interaction networks that we will gain a basic understanding on biological processes. Since the early 50s, this realization has led to the birth and rise of "structural biology", the science of elucidating structures of biomolecules at atomic scale and relating them to their functions.

Classical structural biology is mainly defined by two experimental techniques, X-Ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy, through which atomic structures of tens of thousands of biomolecules, mainly proteins, and thousands of biomolecular complexes have been solved in an accurate manner. These methods have helped immensely in the discovery of the structural world of many biomolecular complexes. Still, as a field, structural biology is far from keeping pace with the speed at which biological data are being generated in other disciplines, such as biochemistry. As an example, the current number of known atomic structures of macromolecular complexes is depicted to be considerably smaller than the documented protein-protein interactions. Unfortunately, technical limitations of classical structural biology techniques and/or the nature of the biomolecules under study hamper to close this gap in a rapid manner. As a rescue strategy, structural biologists often resort to different types of biochemical and biophysical experiments that can quickly provide low-resolution information on macromolecular complexes, even for challenging cases. Most of the time, however, the collected data are rather sparse and/or contain limited information compared to that obtained from classical structural methods. These limitations call for integrative computational tools that can, by using a physics-based model, judiciously combine and accurately translate sparse experimental data into structural information. One such tool is macromolecular docking, defined as the process of building a molecular complex from its known individual components.

The generic idea behind docking approaches can be comprehended easily via the following analogy (David Goodsell, The Machinery of Life, p.10):

"*Constructing molecular complexes is much like trying to build machines with Lego blocks: you may build a variety of different objects, but the final form is shaped and limited by the* ***shape*** *and* ***connections*** *of the underlying units.*"

Current "dockers" have exploited this analogy in two ways: They have either constructed algorithms based on limitations imposed by ***shape*** (shape complementarity methods) or ***connections*** (optimization methods that minimize a target energy function to satisfy connections –restraints- between supplied molecules). Albeit different in their approaches, both methods are confronted with similar challenges when building multi-component assemblies, addressing large conformational changes upon binding or selecting the "correct" solution among generated models. The work described in this thesis focuses on extending the

capabilities of our *in-house* docking program **HADDOCK** by developing new protocols and incorporating new types of experimental information for being able to tackle those challenges.

In **Chapter 1**, I provide a comprehensive introduction on the concept of integrative modeling. For that, I first introduce the sources of low-resolution information and then I describe various methods to integrate that information into the modeling process. I also depict the major challenges that the integrative modeling field is faced with: **(1)** Modeling large assemblies, i.e. dealing with multiple molecules simultaneously, **(2)** modeling dynamic molecular complexes, i.e. addressing large conformational changes upon binding, **(3)** constructing accurate scoring functions for fishing out the biologically relevant solution(s) among the generated pool of conformers, and **(4)** dealing with the degeneracy and ambiguity of methodological advances regarding those challenges taken from literature. I end this chapter by introducing our data-driven docking approach HADDOCK, which can integrate various sources of information to drive the modeling of biomolecular complexes. In the following **Chapters 2-5**, I position HADDOCK within the described integrative approaches, by presenting the protocols I developed during my PhD studies to cope with the depicted challenges (**Chapter 2** refers to challenge (1), **Chapter 3** to (2), **Chapter 4** to **(3)** and **Chapter 5** to (4)).

In **Chapter 2** I demonstrate that HADDOCK is able to handle multiple molecules and allows to impose different types of symmetries during docking. All of these options are implemented into a novel web-interface (*the Multi-body interface*) of HADDOCK, which allows the user to dock up to 6 biomolecules simultaneously, offers inclusion of experimental and/or bioinformatics data and supports several types of cyclic and dihedral symmetries during docking. The performance of the web-server is tested on a benchmark of six cases, containing five symmetric homo-oligomeric protein complexes and one symmetric protein-DNA complex. The quality of the modeled multimeric assemblies reveal that, in the presence of either bioinformatics and/or experimental data, HADDOCK is able to generate high-ranking near-native solutions (having an interface-RMSD range of 0.8 – 1.8Å) for all cases, demonstrating its ability to model symmetric multi-component assemblies.

In order to address the challenge of modeling binding-induced large-scale conformational changes, in **Chapter 3**, I describe a novel *Flexible Multi-domain* (FMD) docking protocol. FMD aims at modeling both large-scale domain (hinge) motions and small- to medium-scale interfacial rearrangements: this is achieved by cutting the flexible molecule at a hinge region predicted by an elastic network model, and then by performing a simultaneous docking of all sub-parts/domains making use of HADDOCK's multi-body docking ability, presented in **Chapter 2**. The cut domains are kept together by connectivity restraints. The performance of this unprecedented range of conformational changes, from 1.5 to 19.5Å. In general, this FMD protocol could reproduce at least a near-native solution (interface-RMSD $\leq$ 4Å) for each case and rank these at the top. These results demonstrate that this protocol is poised to model conformational changes as large as 20Å! Finally, I show that the cumulative sum of eigenvalues obtained from the elastic network is indicative to predict the extent of the conformational change to be expected.

In **Chapter 4**, I demonstrate that the scoring problem can be eased by incorporating information-based terms into the scoring function. For this purpose, I integrate low-resolution shape information obtained from either Ion Mobility Mass Spectrometry (IM-MS) or Small Angle X-Ray Scattering (SAXS) experiments into HADDOCK's scoring function and systematically assess the strengths and weaknesses of IM-MS- and SAXS-based scoring. For that, I make use of a large docking decoy set composed of 138 heterodimers generated by running HADDOCK in *ab initio* mode. The statistics calculated on this large decoy set suggests that IM-MS data are of too low resolution for selecting correct models, while scoring with SAXS data leads to a significant performance improvement. The performance of SAXS scoring however depends on the shape and arrangement of the complex and its constituents.

In **Chapter 5**, I present two application examples of integrative modeling using HADDOCK. In the first one, I combine NMR chemical shift perturbations and mutagenesis data in order to predict the topology of the complex between the deubiquitination enzyme Josephin and a di-ubiquitin chain. The results of this integrative modeling suggest that Josephin imposes a preference for the cleavage of K48-linked di-ubiquitin chains, which is validated by biochemical experiments. In the second example, I model the transfer of the histone heterodimer H3-H4 from one histone chaperone, Asf1, to the other one, p48, based on Electron Paramagnetic Resonance data that suggests large structural rearrangements upon complexation and SAXS data, which provides insight on the globular assembly of the complex. This transfer takes place during the initiation of nucleosome assembly. The generated integrative models suggest putative inter-molecular interactions that can be tested by further mutagenesis experiments.

In the final chapter of my thesis, **Chapter 6**, I give a perspective on future directions of HADDOCK. These include incorporating low-resolution shape data (coming from SAXS, cryo-EM) into sampling, implementing coarse-grained representations into HADDOCK in order to allow the modelling of even larger assemblies, and adapting HADDOCK's energy functions in such way that it will enable the modelling of membrane systems. At the end of this chapter, I provide a proof-of-principle example to underpin the fact that the accuracy of integrative models not only depends on the presented methodological advances and future directions, but also on the quality and the distribution of the data supplied. This example nicely pinpoints the fact that integrative modeling should not be seen as an end point, but rather a starting point for generating new hypothesis to be tested experimentally.

**In conclusion,** the recent and ongoing developments in integrative modeling, together with the ones presented in this thesis, are important milestones in the modeling of protein-protein interactions. This should lead in the next decade to a significant growth in the field of structural computational biology and pave the route to increased accuracy, coverage and resolution of integrative models. Comprehending the underlying mechanistic of biomolecular function is not a simple process; computational structural biologists and experimentalists will therefore have to work in a collaborative and synergistic manner, in order to make significant advances in this field.

# Samenvatting

Alle levensprocessen die we dagelijks om ons heen zien, voelen, ondergaan of aanschouwen worden in wezen gereguleerd op de *nanoschaal* ($10^{-9}$ m). Of nog preciezer, de biologische functionaliteiten die we op de *macroschaal* waarnemen zijn een optelsom van gearrangeerde communicaties tussen biomoleculen, voornamelijk eiwitten, op de *nanoschaal*. Door dus inzicht te krijgen in de *nanoschaal* wereld van biomoleculen en hun netwerk van interacties te ontleden, kunnen we ons fundamentele begrip van biologische processen vergroten. Dit bewustzijn heeft vanaf het begin van de jaren vijftig geleid tot het ontstaan en groeien van de "structuur biologie", de wetenschap die zich bezighoudt met het ophelderen van de atomaire structuur van biomoleculen om deze vervolgens te relateren aan hun biologische functies.

De klassieke structuur biologie wordt hoofdzakelijk bepaald door twee experimentele technieken waarmee de atomaire structuren van tienduizenden biomoleculen, voornamelijk eiwitten, en duizenden biomoleculaire complexen zorgvuldig bepaald zijn, namelijk X-ray kristallografie en *Nuclear Magnetic Resonance* (NMR). Deze methoden hebben een enorme bijdrage geleverd bij het ontrafelen van de vouwingswereld van veel biomoleculaire complexen. Desalniettemin kan het wetenschapsgebied van de structuur biologie met geen mogelijkheid de snelheid bijbenen waarmee biologische gegevens worden gegenereerd in andere disciplines, zoals de biochemie. Het huidige aantal opgehelderde atomaire structuren van macromoleculaire complexen is bijvoorbeeld vele malen kleiner dan alle in de literatuur beschreven eiwit-eiwit interacties. Technische beperkingen van de klassieke structuur biologie technieken en/of de aard van de onderzochte biomoleculen maken het helaas lastig deze kloof in de nabije toekomst te dichten.

Als uitweg voeren structuur biologen vaak verschillende biochemische en biofysische experimenten uit, die zelfs in lastige gevallen redelijk snel lage resolutie informatie over macromoleculaire complexen kunnen verschaffen. In de meeste gevallen is de informatie die de experimenten leveren echter schaars en/of beperkt vergeleken met de klassieke structuur biologie methoden. Dit zorgt voor een vraag naar computationele methoden die op een intelligente manier deze schaarse experimentele gegevens kunnen integreren, combineren en op een juiste manier kunnen vertalen naar structuur informatie: *integratief modeleren*. Een voorbeeld hiervan is macromoleculair dokken, een methode waarbij een complex tussen moleculen wordt opgebouwd uit de reeds bekende individuele componenten.

De volgende vergelijking legt op een simpele manier het generieke idee achter dokken uit (David Goodsell, The Machinery of Life, p.10):

"Het construeren van moleculaire complexen is als het bouwen van machines met lego blokken: je kan verschillende objecten bouwen, maar het uiteindelijke resultaat wordt bepaald en gelimiteerd door de **vorm** en mogelijke **verbindingen** van de gebruikte onderdelen."

De huidige "dokkers" hebben deze analogie op twee manieren uitgewerkt: ze hebben algoritmes ontwikkeld op basis van de voorwaarden die door de ***vorm*** worden opgelegd (complementaire vorm methoden) óf door de ***verbindingen*** (methoden die een energiefunctie minimaliseren om te voldoen aan de verbindingen -*restraints*- tussen de relevante moleculen). Hoewel de twee benaderingen verschillend zijn, worden beide methoden met dezelfde uitdagingen geconfronteerd als het gaat om het opbouwen van complexen die bestaan uit meerdere componenten, het verwerken van grote conformationele veranderingen tijdens het vormen van het complex of het selecteren van de "juiste" oplossing uit de genereerde modellen. Het werk beschreven in dit proefschrift heeft als doel het uitbreiden van de mogelijkheden van ons *in-house* dokprogramma HADDOCK, door nieuwe protocollen te ontwikkelen en nieuwe soorten experimentele gegevens toe te voegen, om de hiervoor beschreven uidagingen aan te gaan.

In **Hoofdstuk 1** leg ik op een begrijpelijke manier het concept *integratief modeleren* uit. Hiertoe introduceer ik eerst de technieken die lage resolutie informatie kunnen verschaffen en vervolgens beschrijf ik verschillende methoden die deze informatie kunnen integreren in het modeleringsproces. Daarnaast geef ik aan wat de belangrijkste uitdagingen zijn waar het onderzoeksveld van het integratief modeleren mee geconfronteerd wordt: **(1)** Het modeleren van grote assemblages, ofwel het werken met meerdere moleculen tegelijkertijd, **(2)** het modeleren van dynamische moleculaire complexen, ofwel het verwerken van grote conformationele veranderingen door het binden van de moleculen, **(3)** het ontwerpen van accurate scoringsfuncties waarmee de biologisch relevante oplossing(en) uit de populatie van gegenereerde modellen geselecteerd kan (kunnen) worden en **(4)** het kunnen werken met gedegenereerde en ambigue startgegevens. Na dit gedeelte, schets ik de meest relevante in de literatuur beschreven methodologische ontwikkelingen ten aanzien gedreven dokprogramma HADDOCK dat informatie van verschillende experimentele origine kan integreren, om zo het modeleren van biomoleculaire complexen te sturen. In de hoofdstukken die volgen (**Hoofdstukken 2-5**) plaats ik HADDOCK ten opzichte van de beschreven integratieve methoden en beschrijf ik de protocollen die ik tijdens mijn promotieonderzoek heb ontwikkeld met betrekking tot de hierboven genoemde uitdagingen (**Hoofdstuk 2** verwijst naar de **1$^{e}$** uitdaging, **Hoofdstuk 3** naar de **2$^{e}$**, **Hoofdstuk 4** naar de **3$^{e}$** en **Hoofdstuk 5** naar de **4$^{e}$**).

In **Hoofdstuk 2** laat ik zien dat HADDOCK meerdere moleculen kan hanteren en in staat is tijdens het dokken verschillende soorten symmetrie op te leggen. Al deze opties zijn opgenomen in een nieuwe HADDOCK web-interface (de *Multi-body interface*) waar de gebruiker de mogelijkheid heeft tot 6 moleculen tegelijkertijd te dokken, experimentele gegevens of bioinformatische voorspellingen kan gebruiken en meerdere soorten cyclische en dihedrale symmetrieën tijdens het dokken kan toepassen. Om de prestatie van de web-server te bepalen is deze getest met een *benchmark* bestaande uit vijf symmetrische homo-oligomere eiwitten en één symmetrisch eiwit-DNA complex. De kwaliteit van de gemodelleerde multi-componente assemblages laat zien dat wanneer bioinformatische voorspellingen en/of experimentele gegevens gebruikt worden, HADDOCK in alle gevallen *near-native* oplossingen (met een interface RMSD tussen 0.8 – 1.8Å) kan genereren die

hoog gerangschikt worden en laat dus zien dat HADDOCK in staat is symmetrische multi-component assemblages te modeleren.

In **Hoofdstuk 3** beschrijf ik een nieuw *Flexible Multi-domain* (FMD) dokprotocol waarmee ik grote conformationele veranderingen kan modeleren die door het binden geïnduceerd worden. FMD heeft als doel het modeleren van zowel grote domein (scharnier) bewegingen als ook kleine tot middelgrote herschikkingen aan de interface: dit wordt tot stand gebracht door het flexibele molecuul in een scharniergebied, dat door een elastisch netwerk model voorspeld kan worden, door te knippen en vervolgens gebruik te maken van het multi-component dokprotocol in HADDOCK beschreven in **Hoofdstuk 2**, om tegelijkertijd alle onderdelen/domeinen te dokken. De losgeknipte domeinen worden bij elkaar gehouden door connectie *restraints*. De prestatie van deze methode werd getest met een *benchmark* van elf complexen die een ongeëvenaarde omvang van conformationele veranderingen vertegenwoordigen: 1.5 – 19.5Å. Samengevat kan dit FMD protocol op zijn minst *near-native* oplossingen (interface-RMSD $\leq$ 4Å) voor ieder complex genereren en deze bovenaan rangschikken. Deze resultaten geven aan dat dit protocol klaar is om de door het elastische netwerk gegeven cumulatieve som van de eigenwaardes indicatief is voor de grootte van de te verwachtte conformationele verandering.

In **Hoofdstuk 4** laat ik zien dat het scoringsprobleem vereenvoudigd kan worden door op informatie gebaseerde termen aan de scoringsfunctie toe te voegen. Hiertoe heb ik lage resolutie vorminformatie, verkregen van *Ion Mobility Mass Spectrometry* (IM-MS) of *Small Angle X-ray Scattering* (SAXS) experimenten, in de HADDOCK scoringsfunctie geïntegreerd en heb ik systematisch de sterke en zwakke punten van de op IM-MS en SAXS gebaseerde scores getest. Hiervoor heb ik gebruik gemaakt van een set van dok *decoys* bestaande uit 138 heterodimeren die gegenereerd zijn door HADDOCK in de *ab-initio* stand te laten lopen. De statistieken bepaald met deze grote *decoy* set laten zien dat IM-MS gegevens vaak een te lage resolutie hebben om de juiste oplossingen te kunnen selecteren, maar dat SAXS gegevens daarentegen wel voor een significantie verbetering van de scoringsfunctie zorgen. De kracht van de SAXS score hangt echter wel af van de vorm en organisatie van het complex en zijn componenten.

In **Hoofdstuk 5** presenteer ik twee voorbeelden van toepassingen van integratief modeleren met HADDOCK. In het eerste geval combineer ik veranderingen van NMR chemische verschuivingen en mutagenese gegevens om de topologie van het complex tussen het de-ubiquitinatie enzym Josephin en een di-ubiquitine keten. De resultaten van dit integratieve modeleringsvoorbeeld suggereren dat Josephin een voorkeur heeft voor het splitsen van K48-gelinkte di-ubiquitine ketens, hetgeen vervolgens door biochemische experimenten gevalideerd is. In het tweede voorbeeld modeleer ik de overdracht van de histon H3-H4 heterodimeer van de ene histonchaperonne, Asf1, naar de andere, p48, gebruikmakend van *Electron Paramagnetic Resonance* (EPR) en SAXS gegevens. De EPR gegevens wijzen erop dat er grote conformationele verandering plaatsvinden tijdens de complexformatie en de SAXS data geven een inzicht in de algehele organisatie van het complex. Deze chaperonne overdracht vindt plaats tijdens de initiatie van nucleosoom assemblage. Het gegeneerde integratieve model wijst op mogelijke contacten tussen de

moleculen, die nu kunnen worden geverifieerd met behulp van mutagenese experimenten.

In het laatste hoofdstuk van mijn proefschrift, **Hoofdstuk 6**, geef ik mijn visie op de wegen die HADDOCK in de toekomst in zal slaan. Dit omvat het toevoegen van lage resolutie vorm gegevens (afkomstig van SAXS, cryo-EM) in de samplingroutine, het implementeren van *coarse-grained* representaties in HADDOCK om nog grotere assemblages te kunnen modeleren en het aanpassen van de energiefuncties van HADDOCK zodat ook het modeleren van membraan systemen mogelijk zal worden. Aan het eind van dit hoofdstuk bespreek ik een *proof-of-principle* voorbeeld om kracht bij het feit te zetten dat de juistheid van het integratief modeleren niet alleen bepaald wordt door de hier gepresenteerde methodologische vooruitgangen en toekomstige ontwikkelingen, maar ook door de kwaliteit en de distributie van de experimentele gegevens. Dit voorbeeld geeft mooi aan dat het integratief modeleren niet als een einddoel op zich gezien moet worden, maar meer als een uitgangspunt om nieuwe hypotheses te kunnen generen die vervolgens experimenteel getest kunnen worden.

Concluderend, de recente en huidige ontwikkelingen in het integratief modeleren, samen met degenen die in dit proefschrift beschreven zijn, zijn belangrijke mijlpalen in het modeleren van eiwit-eiwit interacties. In het volgende decennium zouden deze moeten leiden tot een significante groei van het veld van de computationele structuur biologie en de weg moeten vrij maken voor een betere precisie, dekking en resolutie van de integratieve modellen. Het begrijpen van de onderliggende mechanismes van de biologische functie is echter geen simpel proces; computationele structuur biologen zullen hiertoe op een synergetische manier samen moeten werken met experimentalisten om significante vooruitgangen in dit onderzoeksgebied te kunnen boeken.

# Acknowledgments

Ten years back from now, I started studying Chemical Engineering. The first year was harsh: I was struggling to keep up with the rest of the class, but didn't know how and I was somehow feeling lost. At the end of freshmen year I met Elif, who became my very best friend later on. Regardless of many great moments we've had together, we've also learned a lot from each other. I learned "live the moment philosophy" means to comprehend the lecture during the time course of the lecture(:p), prioritization matters and most importantly going into the depth of a subject is what makes the real difference. With my deepest honesty I can say that during my early university years, my scientific thinking was initially shaped by Elif!

In my junior year, I took a *Thermodynamics* class from Türkan Haliloğlu, where for the first time I was not only interested into a course but also into the philosophy behind it. I was amazed by how I could relate the course content to our daily lives, and how great that felt. I was also astonished by Türkan's enthusiasm and her big smile she always had during the classes. By then, I knew I wanted to do my graduation project with Türkan. When I consulted her about this, she invited me to her group with great sincerity and opened the doors of computational biology field to me. I found out that I loved working in this field so much that I wanted to continue with the master's program of Chemical Engineering department, which allowed me to work further with Türkan and immerse in the grounds of computational biology.

During my masters, with support of Türkan I had the opportunity to visit National Cancer Institute in the US and work with Ruth Nussinov, who is the most energetic and enthusiastic scientist I've met. If I could produce anything solid during my masters, that's thanks to those assets I've inherited from them.

Coming to my PhD, which stands as a major milestone in my life not only for my career but also in the means of leaving my hometown and standing on my feet alone, I believe I've made one of my luckiest and best choices in life. And that choice's name is Alexandre. In June 2008, he visited our lab in Istanbul and gave a presentation about docking and HADDOCK. When he finished, I was sure that I have just seen what I want to do during my PhD. While trying to suppress my excitement, I approached him and introduced myself, which started the chain of events leading me to write this Acknowledgements.

Dear Alex, when people ask me about you and your skills as a supervisor, I have been always telling them proudly "He is just GREAT!". You are a keen, clever, enthusiastic, efficient, honest and fun scientist: such a genuine character! I don't remember any single day I haven't seen you smiling, or making jokes. Whenever I needed, I could receive your consultancy and having such a supervisor was one of the major motivations during my PhD. It was also big pleasure to travel with you for

conferences/workshops (especially to Istanbul!). You've been a great role model and you set the standards very high for my future career that I'll try very hard to keep up with.

Now the best lab crew; Computational Structural Biology (CSB) group:

Adrien, my dearest office-mate, my paranymph, who helped me seeing the optimistic side of the incidents (both scientific and non-scientific). It's with your nice efforts that I adapted quicker to cold and rainy -um sorry refreshing- weather and everything else :) In the office we had so much fun together, I'll mostly miss our "air high-fives" and the moments we shared our teaching experiences with the others (at those moments even you can get mean :)). One last remark; I am so proud that with our collaborative efforts, we could make our office the warmest one in the whole lab!

Panos, my dearest kanka, my fellow PhD student, the number one party-boy of the lab. During the years whenever we worked in a collaborative manner, especially for CAPRI, the result was always amazing! That's why we'll send our students to each other when we'll have a lab in the future, right? I am so grateful that I was your paranymph during your defense! I was really proud of us (you, me and João), we made the best (looking) team ever! One last thing, don't forget to thank me when you receive the Nobel Prize, I'll be watching!

João, my Portuguese kanka! I don't know how you do it, but it's impossible not to like you: you are nice, fun, handsome and a true gentleman. When I was living in Istanbul, I used to walk home with friends to discuss about life-issues. Many years later, I experienced the same thing with you, which reminded me of home! Also, this comes both to Steven and to you: when is the dinner you promised guys, you just cannot postpone it anymore!!

Dear Marc, it's amazing that whatever you do, you do the best! You are the only protein-DNA docking expert in the lab and also you make the most esthetic designs. I should here admit that whenever I prepare a poster I was trying to imitate your style, and apparently this was such a clever idea as it made me win two poster prizes, not bad, huh?:)

Mikmiiik!! My French fella! You are definitely the funniest person of our group (sorry Panos!). Kanka, your sarcastic sense of humor has changed the atmosphere of the lab, such that with you everyday in the lab was like a party. So, I feel very sorry since you left the lab, but as one great Turkish philosopher pointed out: "What can I do, sometimessss?".

Christophe, why did you leave, why, why? Ok, I love Adrien (see above) but know anymore how is Marseilles doing against Lyon, how is the current French politics, also I don't fancy to drink cola in the afternoon (maybe this is not so bad?). But, I know that you'll be doing fine with Chantel, you are so lucky to have a girlfriend cooking so tasty dishes!

Klaartje, I am so glad that I met a woman like you: you are an excellent scientist, caring supervisor and the mother of my favorite children, Gael and Elliot☺. Many thanks for making me a part of your family atmosphere; that felt like home! Also, thanks really a lot for all of your efforts to translate my summary, that meant a lot to me and I really appreciate it!!

Sjoerd, thanks for sharing your deep knowledge in HADDOCK and structural biology! During the assembly of my thesis, I had a final chance to critically read our MCP paper (Chapter 2). I have to note that I'm so proud of this paper, so many thanks to you Sjoerd, Adrien, Panos and Marc, for helping me to get out such a comprehensive work! Gydo, the newest-brilliant member of our team! After me you'll be the next prospective "data-integrate-or", I count on you, don't let me down!

Dearest Charleen, don't make me miss our conversations, stay in touch! I would be happier if you would have been my student, but I am sure that you are also sort of doing fine with João and Marc:p Iva and Arne, I hope that I could be a good supervisor for you. From my side, I really enjoyed working and discovering new fields with you. Good luck with the rest of your careers, I am sure that you'll be doing just fine! Koen, you managed Panos very well, congratulations for that!

Rolf, I still couldn't get how you know everything about everything! Maybe without noticing, you've helped me to improve my research: thanks to your to-the-point questions! Gert, thanks for sharing nice borrel memories, Barbara and Johan thank you for helping me whenever I needed it.

Tobias, thanks a lot for making me to understand how to interpret the SAXS data! Markus, Tessa and also dear seniors: Hans, Hugo, Mark, Tsjerk, Nuno, Dirk and the rest of the NMR lab, it was really great to have you around (Hans, Markus and Mark Danielopulos, especially during borrels:)).

Dear Marietje, it's been a while, but be sure that your spirit and your lovely smile has been always with us! I hope that once we can sail together in Turkey, maybe even during this summer?:)

Lovely Maryam, hold on to your beautiful niceness, as that is what makes you special. Maryam, Mohammed, Deepak and Rama, you made me experience the "eastern essence" when I miss home, thanks for making me feel less homesick!

Sweet Annalisa, my sister! I'm indebted to you for making me feel home during my early years in the Netherlands. Without your support and friendship, here, I would have felt much colder. Carles, after the meeting in Saint Feliu, I couldn't drink cava for more than a year and you know why! Such a fantastic memory! Seren, thank you so much for helping me to organize the HADDOCK workshop in Istanbul. Also, thanks to all Polymer Research Center members for their constant support!

Ultimately it's the turn of my Turkish-gang living (or lived) in the Netherlands: My biricik kanka Ali Alejandro, mega-mouth Ata Can, amazing cook Erman, brilliant Elif, sincere Dilek, messy-but-birokadarda-funny Ahmet, true

geologist Ali Ö., beautiful Nazem, hospitable kanka Ceylan, motorcyclist Sinan, sweet Dicle, big chef Yunus, genial Bahar, stylish Sevgi and diplomat Efe basgan. With smaller or bigger contributions, you all made my life enjoyable in the Netherlands, endless gratitude for that!

Erman, maybe I would have anyhow received a PhD degree, but it wouldn't be the same, it wouldn't be this good without your support! To express my appreciation, I hereby declare that -by using my prospective degree- I entitle you as an honorary PhD in Computational Structural Biology! Alex still doesn't know about this, but I am sure that he'll be fine with that:).

## Epilogue

The first book I read was a gift from my father. It's called the "*Little Black Fish*" written by the Persian writer Samed Behrengi. Any child who read the following part of this book could never stay the same:

"*I want to go see where the stream ends. You know, Mother, I've been wondering where the end of the stream is. I haven't been able to think about anything else. I didn't sleep a wink all night. I want to know what's happening in other places. I want to know if life is simply for circling around in a small place until you become old and nothing else, or is there another way to live in the world?*"

I have become a scientist and left my home country to see where the stream ends. If I could dare to start this long journey, this is all thanks to the endless courage I have received from my beloved family, my mother **Sevgi** and my father **Avni**. This thesis is dedicated to them.

**Ezgi**

A

# Publications

## HADDOCKing-related:

- **E. Karaca**, A.M.J.J. Bonvin, On the usefulness of Ion Mobility Mass Spectrometry and SAXS data in scoring docking decoys (*submitted for publication*).
- **E. Karaca**, A.M.J.J. Bonvin, Advances in Integrative Modeling of Biomolecular Complexes, *Methods, 2012, in press.*
- J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, **E. Karaca**, A.S.J. Melquiond and A.M.J.J. Bonvin, Clustering biomolecular complexes by residue contacts similarity, Proteins: Struc. Funct. & Bioinformatics 80(7):1810-7, 2012
- A.S.J. Melquiond, **E. Karaca**, P.L. Kastritis, A.M.J.J Bonvin, Next challenges in protein–protein docking: from proteome to interactome and beyond, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2: 642-651, 2012.
- C. Schmitz, A.S.J. Melquiond, S.J. de Vries, **E. Karaca**, M. van Dijk, P.L. Kastritis and A.M.J.J. Bonvin, Protein-protein docking with HADDOCK, In: *NMR of Biomolecules: Towards Mechanistic Systems Biology*, Edited by I. Bertini, K.S. McGreevy and G. Parigi, Wiley-VCH, 512–535, 2012.
- **E. Karaca**, A.M.J.J. Bonvin, A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes, *Structure*, 19-4, 555-565, 2011.
- **E. Karaca**[*], A.S.J. Melquiond[*], S.J. de Vries[*], P.L. Kastritis, A.M.J.J. Bonvin, Building Macromolecular Assemblies by Information-driven Docking, *Molecular & Cellular Proteomics*, 9-8, 1784-1794, 2010.
- S.J. de Vries, A.S.J. Melquiond, P.L. Kastritis, **E. Karaca**, A. Bordogna, M. van Dijk, J.P. Rodrigues, A.M.J.J. Bonvin, Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions, *Proteins: Structure, Function, and Bioinformatics*, 78-15, 3242-3249, 2010.
- G. Nicastro, S.V. Todi, **E. Karaca**, A.M.J.J. Bonvin, H.L. Paulson, A. Pastore, Understanding the role of the Josephin domain in the PolyUb binding and cleavage properties of ataxin-3, *PloS one*, 5-8, e12430, 2010.

## Others:

- **E. Karaca**, M. Tozluoglu, R. Nussinov, T. Haliloğlu, Alternative Allosteric Mechanisms Can Regulate the Substrate and E2 in SUMO Conjugation, *Journal of Molecular Biology*, 406-4, 620-630, 2011.
- M. Tozluoğlu, **E. Karaca**, R. Nussinov, T. Haliloğlu, A Mechanistic View of the Role of E3 in Sumoylation, *PLoS computational biology*, 6-8, e1000913, 2010.
- M. Tozluoğlu[*], **E. Karaca**[*], T. Haliloğlu, R. Nussinov, Cataloging and organizing p73 interactions in cell cycle arrest and apoptosis, *Nucleic acids research*, 36-15, 5033-5049, 2008.

[*]Equal Correspondence

# Curriculum Vitae

Ezgi Karaca was born in Istanbul, Turkey, on July 29, 1983. She graduated from the science section of the Cağaloğlu Anatolian High School in 2001. In the same year she entered Boğaziçi University and enrolled in Turkey's top ranking Chemical Engineering program. During the final year of her bachelor studies (2006), she joined Polymer Research Center (PRC) and started to work with Prof. Dr. Türkan Haliloğlu. In 2008, she started her masters in the same program and continued her research under the umbrella of PRC. Throughout her masters, she visited the laboratory of Prof. Dr. Ruth Nussinov in National Cancer Institute of USA during two summers. In those years, her work was focused on the p73's tumor suppression network and understanding the underlying mechanistic of sumoylation cascade. Following her studies in Turkey, she started her PhD in October 2008 in the Computational Structural Biology Group of Prof. Dr. Alexandre Bonvin to carry out the PhD project described in this thesis. During the last year of her PhD she was entitled "the PhD of the year" of Bijvoet Center for Biomolecular Research. She will defend her thesis on 6$^{th}$ of February 2013.

C