Methods 59 (2013) 372–381

Contents lists available at SciVerse ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Advances in integrative modeling of biomolecular complexes

Ezgi Karaca, Alexandre M.J.J. Bonvin *

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

ARTICLE INFO

Article history:
Available online 23 December 2012
Communicated by Peter Schuck

Keywords:
Hybrid methods
Restraints
Information-driven docking
HADDOCK
Computational structural biology

ABSTRACT

High-resolution structural information is needed in order to unveil the underlying mechanistic of biomolecular function. Due to the technical limitations or the nature of the underlying complexes, acquiring atomic resolution information is difficult for many challenging systems, while, often, low-resolution biochemical or biophysical data can still be obtained. To make best use of all the available information and shed light on these challenging systems, integrative computational tools are required that can judiciously combine and accurately translate sparse experimental data into structural information. In this review we discuss the current state of integrative approaches, the challenges they are confronting and the advances made regarding those challenges. Recent developments are underpinned by noteworthy application examples taken from the literature. Within this context, we also position our data-driven docking approach, HADDOCK that can integrate a variety of information sources to drive the modeling of biomolecular complexes. Only a synergistic combination of experiment and modeling will allow us to tackle the challenges of adding the structural dimension to interactomes, shed "atomic" light onto molecular processes and understand the underlying mechanistic of biomolecular function. The current state of integrative approaches indicates that they are poised to take those challenges.

1. Introduction

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 8–20 Å range [3]. But rarely the resolution gets better than 8–6 Å, 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

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.

* Corresponding author. Fax: +31 (0)30 2537623.
E-mail address: a.m.j.j.bonvin@uu.nl (A.M.J.J. Bonvin).

1046-2023
http://dx.doi.org/10.1016/j.ymeth.2012.12.004

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].

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.

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 20–40 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].

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 Å 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].

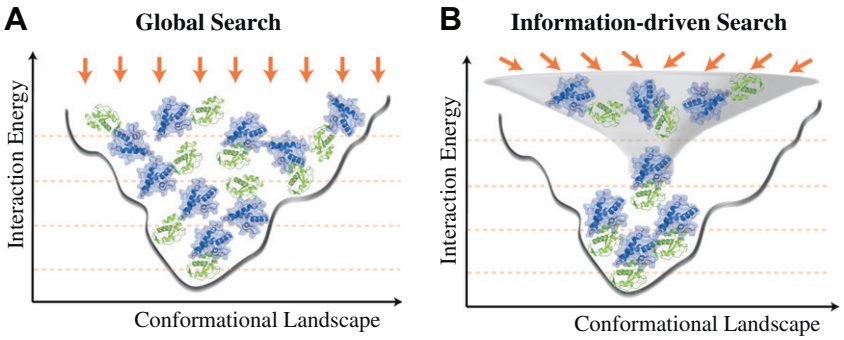

Fig. 1. Global search vs. information-driven search. (A) A global search method performs a thorough exploration of the conformational space resulting in a large number of heterogeneous solutions. (B) An information-driven search method is directed by the data supplied and thus the search is only concentrated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

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], 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 (Fig. 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 (Fig. 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)).

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

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]:

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

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].

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 terms describing steric complementarity, electrostatics, hydrogen bonds, knowledge-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 to use 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 ($\leqslant$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 methods by taking the interfacial shape complementarity into account while fitting 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 promising method 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 ~10 Å away from the known crystal structure, 6 Å 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 the 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 [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 and 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 (Fig. 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].

HADDOCK was originally developed to make use of NMR data, and in particular of chemical shift perturbation data (see Section 2.1.) [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 2 Å, 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

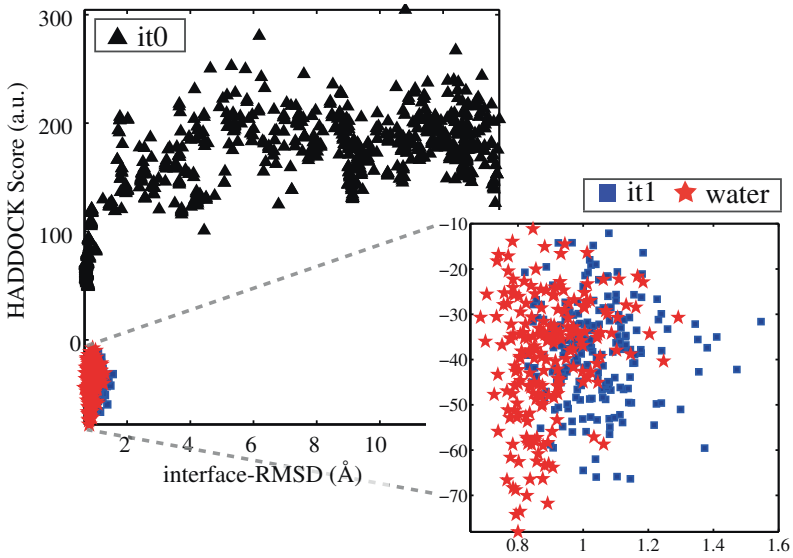

Fig. 2. HADDOCK score as a function of the interface-RMSD for models generated at the various stages of the protocol. Solutions obtained after the initial rigid body energy minimization are indicated by black triangles. These are scored and the top 200–400 hundreds are selected for semi-flexible refinement 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).

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 can impose different combination of restraints at different stages of the docking 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/.

4.1. How does HADDOCK address the challenges of integrative approaches?

HADDOCK is one of the few molecular docking programs able to perform direct docking of generic multibody complexes. It can deal simultaneously with up to six molecules of various natures, including, when needed, cyclic and/or dihedral symmetries [33]. HADDOCK's ability to model symmetric multimers has been benchmarked on five symmetric protein homo-oligomers, including two unbound cases and one symmetric protein–DNA complex. The interaction restraints were obtained from bioinformatics predictions for the protein–protein complexes and from mutagenesis and ethylation interference experiments for the protein-DNA complex. In all cases, native to near-native predictions (with interface-RMSD values for the best model ranging between 0.7 Å and 2.2 Å) were obtained, irrespective of the docking difficulty (Fig. 3) [33].

Next to modeling multimeric macromolecular assemblies, HADDOCK also offers an efficient methodology to deal with large conformational changes upon binding, by performing Flexible Multi-domain Docking (FMD) [95]. The FMD protocol is a *divide-and-conquer* approach: it handles the flexible binding partner as a collection of sub-domains with connectivity restraints between them and uses the multibody docking option of HADDOCK to dock the separated domains simultaneously (Fig. 4). This approach enables modeling of large-scale domain motions at the rigid body docking stage, which is followed by regular flexible refinement, where limited side-chain and backbone conformational changes are addressed. The performance of this protocol was 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 order to focus on the capacity of FMD to model large conformational changes, knowledge of the interface region was assumed. We could demonstrate that by using the same set of restraints FMD significantly outperforms standard two-body docking, generating a top-ranking near-native solution for each case (with interface-RMSDs ranging from 1.1 Å to 4.6 Å). This *divide-and-conquer* approach is thus able to model conformational changes as large as 19.5 Å, something that had never been demonstrated before [8,95].

In HADDOCK, scoring at the end of the initial rigid body docking, *it0*, is very critical, as only a fraction of the generated models (e.g. top 200–400) will be subjected to further flexible refinement. Recently, we generated a docking decoy set by running HADDOCK in *ab initio* mode with only *center-of-mass* restraints on the unbound structures of the Docking Benchmark 4.0 (176 complexes). [130] Analysis of the resulting sets of 10,000 models per complex disclosed that, at the rigid-body stage, HADDOCK could sample at least one near-native solution (interface-RMSD $\leqslant$ 4 Å) in 78% of the cases, whereas in only 38% of them those solutions were selected for further flexible refinement (unpublished data). As mentioned previously, the inaccurate nature of scoring functions can be improved by incorporating experimental data, like for example SAXS scattering curves. In order to test if a SAXS-based scoring could improve the success rate, we added to our scoring function a new term describing the discrepancy between the back-calculated SAXS curve of the model and the experimental (reference) one. As a result, SAXS data could improve the scoring accuracy after rigid body docking (*it0*) by a factor 1.5 (from 38%

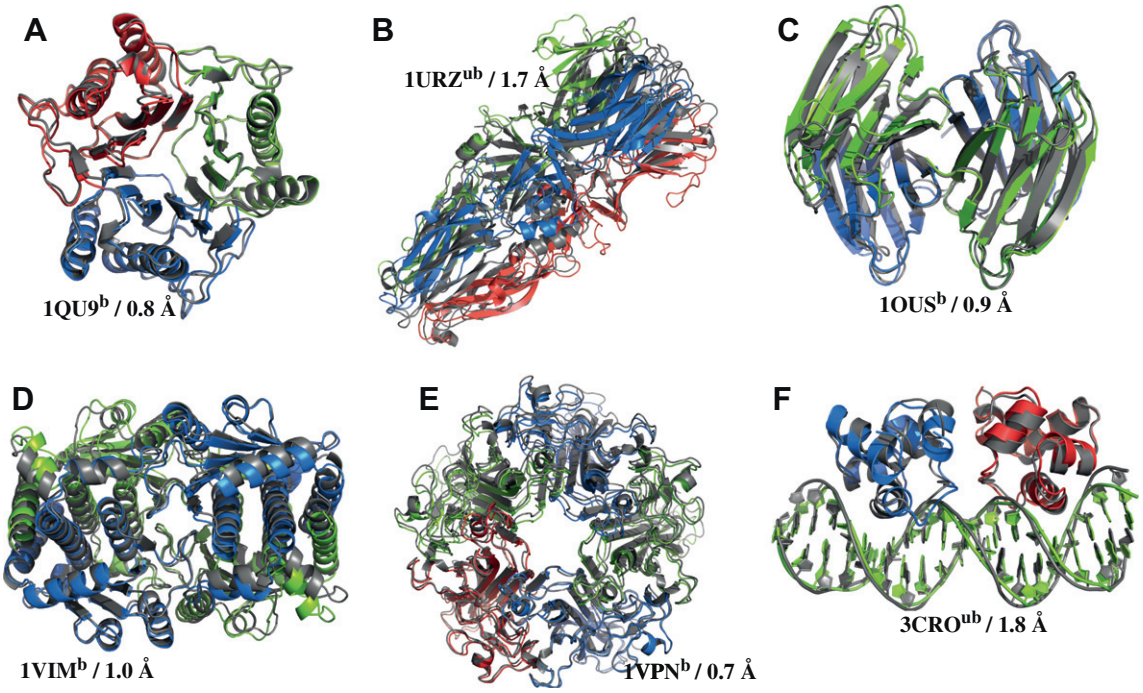

Fig. 3. HADDOCK's performance on symmetric multibody complexes. View of best HADDOCK solutions (with colored monomers) superimposed onto their respective crystal reference structures (shown in dark gray). For each case the type of docking (b:bound, ub: unbound) used to generate that model and its interface-RMSD is reported [33].

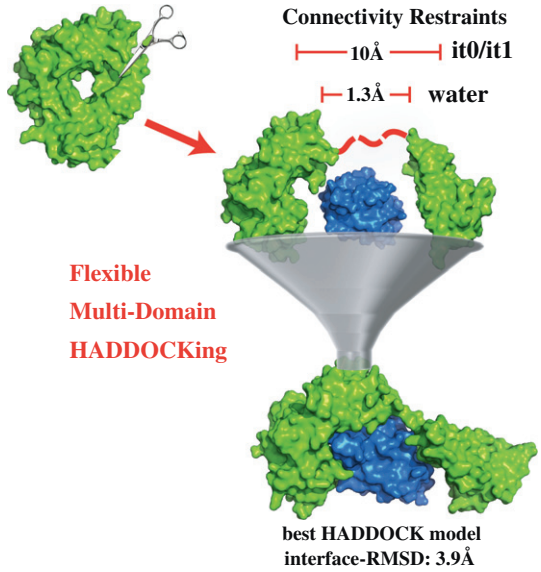

Fig. 4. HADDOCK Flexible Multidomain Docking approach to deal with large conformational changes upon binding. The working principle is illustrated with the interleukin-1 receptor and its antagonist (pdb id: 1ira), experiencing a conformational change of 19.5 Å upon binding [134]. The receptor (green) is cut into domains at its hinge region. A three body docking is then performed on the artificially generated multibody system: two subdomains of the receptor and the ligand (blue). The top ranking model with an interface-RMSD of 3.9 Å is shown [95].

to 57% success rate). This increase is especially pronounced when the shape of one of the complex's constituents is anisotropic (1.6-fold improvement), compared to symmetric ones (1.3-fold improvement). As demonstrated with this example, careful incorporation of experimental information into conventional scoring functions can help to overcome accuracy problems in scoring.

4.1.1. Application examples

Under this section we give two application examples of the use of HADDOCK, in which model building was driven by real experimental data. The first one involves a nice illustration of using ambiguous interaction data in the context of flexible multibody docking, and the second one discusses a degeneracy problem that might be encountered in docking based on cross-linking data.

The first example concerns a deubiquitinating enzyme, Josephin, known to cleave poly-Ubiquitin (Ub) chains. The main question to be addressed was which type of Ub-linkages, K48 or K63, was preferentially cleaved by this enzyme. NMR experiments suggested that 2 Ubiquitin molecules could bind to Josephin simultaneously. Mutational studies combined with NMR chemical shift perturbation data depicted the interaction surface of Ub and Josephin [131]. In particular, two distinct binding sites were identified on Josephin. By inclusion of the stoichiometry, ambiguous interface and di-Ub linkage information concurrently, two linked Ubs were docked onto Josephin following a FMD-like procedure. The resulting models underpinned that Josephin can only properly position K48-linked diUb for cleavage, since the two subunits of K48-linked di-Ub could populate both binding and catalytic sites on Josephin at the same time (Fig. 5A), whereas K63-linked di-Ub chains could not (Fig. 5B). Despite the ambiguity in the interaction data and the limited accuracy of the resulting models, these results were sufficient to propose a hypothesis about preferential cleavage, which was subsequently validated by biochemical experiments (see Fig. 3E in Nicastro et al.'s paper [131]).

In another example, published cross-linked residue pairs of an enzyme-inhibitor complex (collicin-immunity protein in complex with collicin dnase, pdb id: 1ujz) were used to back-calculate the structure [132]. For modeling, the starting monomers were taken from the crystal structure and three inter-monomer distances detected by MS were used to drive the complex formation (Fig. 6A). The docking resulted in a single cluster corresponding to an alternative binding mode (Fig. 6B – gray model). The top ranking solution (Fig. 6B – green model), which did not cluster, was an isolated

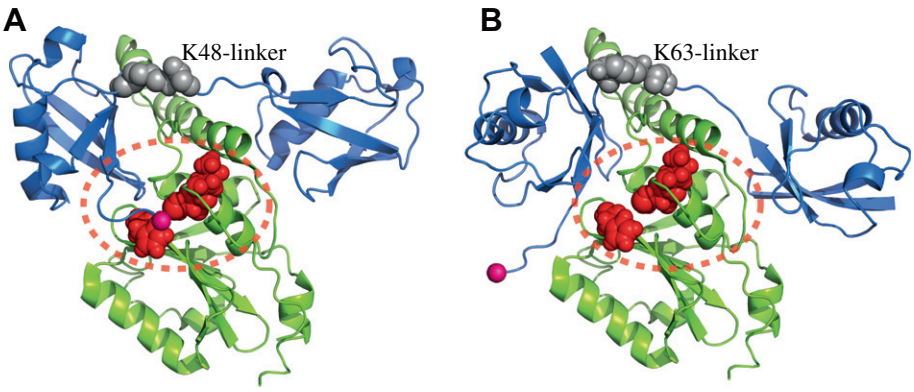

Fig. 5. Information-driven docking with HADDOCK revealed that the topology of Josephin selects preferentiality for K48-linked diUb linkages. Josephin is shown in green cartoon and di-Ubiquitin in marine-blue cartoon with its C-terminus indicated by a magenta sphere. The linked lysines of Ubs are depicted by gray spheres, while the catalytic triad of Josephin is indicated in red and encircled with a dashed ellipse 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 K47-linked poly-ubiquitin chains which was demonstrated experimentally [131].

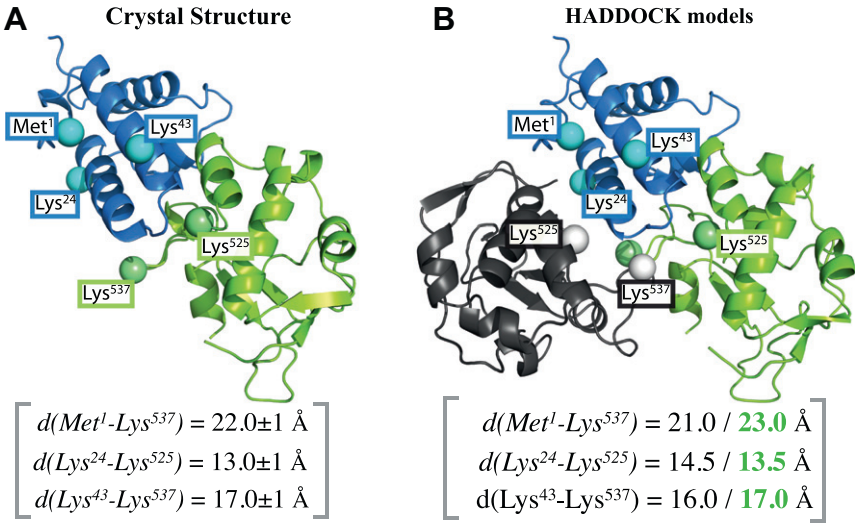

Fig. 6. Docking model's dependency on the quality and degeneracy of the input data. Published cross-linked residue pairs for the enzyme-inhibitor complex (colicin-immunity protein (marine blue) in complex with colicin dnase (green)) were used to drive the docking process. The cross-linked residues are represented in spheres and labeled with their residue id. (A) View of the reference crystal structure (pdb id: 1ujz [135]). The inter-monomer distances used to drive docking are indicated in between brackets. (B) Resulting HADDOCK models: the best ranking cluster, corresponding to an alternative solution, is shown with colicin in black; the isolated (not found in any cluster) top ranking solution, corresponding to the crystal structure, is shown with colicin in green. Both sets of solutions satisfy the cross-link distance restraints. Because of the degeneracy of two cross-links involving the same Lys residues, those three distances are not sufficient to uniquely define the complex.

high accuracy (interface-RMSD $\leqslant$ 1 Å) solution. The alternative binding mode resulted from the fact that two of the inter-monomer distances were sharing the same atom (Lys^{537}, Fig. 4A). To be able to define uniquely the orientation between two surfaces (or two planes) three independent distances are needed [39]. Indeed, when three independent distances that are evenly distributed over the surface of the complex were used, the docking produced a cluster of native solutions with interface-RMSD $\leqslant$ 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 pre-assess both the degeneracy and the quality of any input data before using them during modeling.

5. Conclusions and outlook

A mechanistic understanding of how a cell functions requires complementing interactomes and cellular tomograms by three-dimensional structures of complexes [133]. This daunting task can only be achieved by combining experimentation and computational modeling. Integrative modeling techniques are excellent examples that have been developed to that end. Most of the current macromolecular docking programs, with HADDOCK as one of the pioneers, have stepped into the integrative modeling field.

In this review, we have discussed the current state of integrative approaches and presented the major challenges they are confronting, such as building intricate and dynamic molecular machines, dealing with large conformational changes upon binding or integrating sparse data from various sources. These challenges make it very difficult to properly estimate the accuracy of integrative methods unless their performance has been properly benchmarked. This is a crucial step as it allows to set a proof-of-concept and defines the limits of what is achievable in real applications.

Models originating from integrative approaches should not be seen as an end, but rather as the starting point for new hypothesis

that can be tested experimentally. It will be the synergistic combination of experiment and modeling that will allow us to tackle the challenges of adding the structural dimension to interactomes, shed "atomic" light onto molecular processes and understand the underlying mechanistic of biomolecular function.

Acknowledgements

This work was supported by the Netherlands Organization for Scientific Research (NWO) (VICI grant 700.56.442 to A.M.J.J.B.) and the European Community (FP7 FP7 e-Infrastructure "WeNMR" project, grant number 21301). Support by the national GRID Initiatives of Belgium, France, Italy, Germany, the Netherlands (via the Dutch BiG Grid project), Portugal, Spain, UK, South Africa, Taiwan and the Latin America GRID infrastructure via the Gisela project is acknowledged for the use of web portals, computing and storage facilities. The Josephin project was performed in collaboration with Dr. Annalisa Pastore (MRC National Institute for Medical Research, London UK). We thank Dr. Abdullah Kahraman (ETH) for providing the test case for the cross-link driven docking example discussed in this work.

References

[1] G.C. Lander, H.R. Saibil, E. Nogales, 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., 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., Nucleic Acids Res. 39 (2011) D456–D464.
[4] C. Yang, G. Ji, H. Liu, K. Zhang, G. Liu, F. Sun, et al., 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., 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 20 (2012) 1310–1320.
[7] A. Stein, R. Mosca, P. Aloy, Curr. Opin. Struct. Biol. 21 (2011) 200–208.
[8] A. Melquiond, E. Karaca, P.L. Kastritis, A. Bonvin, WIREs Comput. Mol. Sci. 2 (2011) 642–651.
[9] N.P. Cowieson, B. Kobe, J.L. Martin, Curr. Opin. Struct. Biol. 18 (2008) 617–622.
[10] D. Muradov, B. Kobe, E.N. Dixon, T. Huber, 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, Annu. Rev. Biochem. 77 (2008) 443–477.
[12] A.S.J. Melquiond, A.M.J.J. Bonvin, in: Protein-Protein Complexes: Analysis, Modeling and Drug Design, Imperial College Press, 2010, pp. 183–209.
[13] M.F. Lensink, S.J. Wodak, Proteins Struct. Funct. Bioinf. 78 (2010) 3085–3095.
[14] P.L. Kastritis, A.M.J.J. Bonvin, J. Proteome Res. 9 (2010) 2216–2225.
[15] B. Simon, T. Madl, C.D. Mackereth, M. Nilges, M. Sattler, 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., Cell 149 (2012) 113–123.
[17] T. Madl, F. Gabel, M. Sattler, 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., Science 326 (2009) 1235–1240.
[19] X. Wang, H.-W. Lee, Y. Liu, J.H. Prestegard, J. Struct. Biol. 173 (2011) 515–529.
[20] M.A. McCoy, D.F. Wyss, J. Am. Chem. Soc. 124 (2002) 2104–2105.
[21] S. McKenna, T. Moraes, L. Pastushok, C. Ptak, W. Xiao, L. Spyracopoulos, et al., J. Biol. Chem. 278 (2003) 13151–13158.
[22] K.J. Walters, P.J. Lech, A.M. Goh, Q. Wang, P.M. Howley, 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., Protein Sci. 12 (2003) 811–822.
[24] J.G. Mandell, A.M. Falick, E.A. Komives, Proc. Natl. Acad. Sci. U.S.A 95 (1998) 14705–14710.
[25] G. Pintacuda, G. Otting, J. Am. Chem. Soc. 124 (2002) 372–373.
[26] A. Mohd-Sarip, J.A. van der Knaap, C. Wyman, R. Kanaar, P. Schedl, C.P. Verrijzer, Mol. Cell 24 (2006) 91–100.
[27] T. Clackson, J.A. Wells, Science 267 (1995) 383–386.
[28] W.S. Valdar, J.M. Thornton, J. Mol. Biol. 313 (2001) 399–416.
[29] Q.C. Zhang, D. Petrey, R. Norel, B.H. Honig, Proc. Natl. Acad. Sci. U.S.A 107 (2010) 10896–10901.
[30] F. Pazos, M. Helmer-Citterich, G. Ausiello, A. Valencia, J. Mol. Biol. 271 (1997) 511–523.
[31] I. Kufareva, L. Budagyan, E. Raush, M. Totrov, R. Abagyan, Proteins Struct. Funct. Bioinf. 67 (2007) 400–417.
[32] J. Garcia-Garcia, J. Bonet, E. Guney, O. Fornes, J. Planas, B. Oliva, 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, Mol. Cell. Proteomics 9 (2010) 1784–1794.
[34] G. Otting, K. Wüthrich, Q. Rev. Biophys. 23 (1990) 39–96.
[35] G.M. Clore, Proc. Natl. Acad. Sci. U.S.A 97 (2000) 9021–9025.
[36] E.G. Stein, L.M. Rice, A.T. Brünger, J. Magn. Reson. 124 (1997) 154–164.
[37] M. Trester-Zedlitz, K. Kamada, S.K. Burley, D. Fenyö, B.T. Chait, T.W. Muir, 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., 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., Mol. Cell. Proteomics 9 (2010) 1634–1649.
[40] J. Rappsilber, J. Struct. Biol. 173 (2011) 530–540.
[41] I. Bertini, C. Luchinat, G. Parigi, 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., J. Magn. Reson. 185 (2007) 191–203.
[43] H.J. Steinhoff, Biol. Chem. 385 (2004) 913–920.
[44] G. Otting, Annu. Rev. Biophys. 39 (2010) 387–405.
[45] A.T. Brunger, P. Strop, M. Vrljic, S. Chu, K.R. Weninger, J. Struct. Biol. 173 (2011) 497–505.
[46] A. Cha, G.E. Snyder, P.R. Selvin, F. Bezanilla, Nature 402 (1999) 809–813.
[47] T. Ha, Proc. Natl. Acad. Sci. U.S.A 96 (1999) 9077–9082.
[48] A. Bax, G. Kontaxis, N. Tjandra, Methods Enzym. 339 (2001) 127–174.
[49] R. Brüschweiler, X. Liao, P.E. Wright, Science 268 (1995) 886–889.
[50] A. Grishaev, J. Wu, J. Trewhella, A. Bax, J. Am. Chem. Soc. 127 (2005) 16621–16628.
[51] F. Gabel, B. Simon, M. Nilges, M. Petoukhov, D. Svergun, M. Sattler, J. Biomol. NMR 41 (2008) 199–208.
[52] C.D. Putnam, M. Hammel, G.L. Hura, J.A. Tainer, Q. Rev. Biophys. 40 (2007) 191–285.
[53] J. Kuszewski, A.M. Gronenborn, G.M. Clore, J. Am. Chem. Soc. 121 (1999) 2337–2338.
[54] A. Fotin, Y. Cheng, N. Grigorieff, T. Walz, S.C. Harrison, T. Kirchhausen, Nature 432 (2004) 649–653.
[55] E. Jurneczko, P.E. Barran, Analyst 136 (2011) 20–28.
[56] C. Uetrecht, I.M. Barbu, G.K. Shoemaker, E. van Duijn, A.J. Heck, Nat. Chem. 3 (2010) 126–132.
[57] C. Uetrecht, R.J. Rose, E. van Duijn, K. Lorenzen, A.J. Heck, Chem. Soc. Rev. 39 (2010) 1633–1655.
[58] I. Halperin, B. Ma, H. Wolfson, R. Nussinov, Proteins Struct. Funct. Bioinf. 47 (2002) 409–443.
[59] C. Pons, S. Grosdidier, A. Solernou, L. Pérez-Cano, J. Fernández-Recio, Proteins Struct. Funct. Bioinf. 78 (2010) 95–108.
[60] I.S. Moreira, P.A. Fernandes, M.J. Ramos, J. Comput. Chem. 31 (2010) 317–342.
[61] A. Solernou, J. Fernandez-Recio, J. Phys. Chem. B 115 (2011) 6032–6039.
[62] D.W. Ritchie, G.J. Kemp, Proteins Struct. Funct. Bioinf. 39 (2000) 178–194.
[63] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, I.A. Vakser, 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, Proteins Struct. Funct. Bioinf. 78 (2010) 3197–3204.
[65] C. Dominguez, R. Boelens, A.M.J.J. Bonvin, 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., PLoS Biol. 10 (2012) e1001244.
[67] J.J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, et al., J. Mol. Biol. 331 (2003) 281–299.
[68] W. Braun, N. Go, J. Mol. Biol. 186 (1985) 611–626.
[69] L. Verlet, Phys. Rev. 159 (1967) 98–103.
[70] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. 21 (1953) 1087.
[71] A.T. Brunger, P.D. Adams, L.M. Rice, Prog. Biophys. Mol. Biol. 72 (1999) 135–155.
[72] A.D.J. van Dijk, R. Boelens, A.M.J.J. Bonvin, FEBS J. 272 (2005) 293–312.
[73] Andrew R. Leach, in: Molecular Modelling: Principles and Applications, 2nd ed., Pearson Education, Dorchester, 2001, pp. 368–369.
[74] A.T. Brunger, Nat. Protoc. 2 (2007) 2728–2733.
[75] W.A. Hendrickson, Methods Enzymol. 115 (1985) 252–270.
[76] G.M. Clore, M. Nilges, D.K. Sukumaran, A.T. Brünger, M. Karplus, A.M. Gronenborn, EMBO J. 5 (1986) 2729–2735.
[77] M. Nilges, Curr. Opin. Struct. Biol. 6 (1996) 617–623.
[78] P.D. Adams, N.S. Pannu, R.J. Read, A.T. Brunger, Proc. Natl. Acad. Sci. U.S.A 94 (1997) 5018–5023.
[79] M. Nilges, A.M. Gronenborn, A.T. Brünger, G.M. Clore, Protein Eng. 2 (1988) 27–38.
[80] J.P. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca, A.S. Melquiond, et al., Proteins Struct. Funct. Bioinf. 80 (2012) 1810–1817.
[81] E. Feliu, B. Oliva, Proteins Struct. Funct. Bioinf. 78 (2010) 3376–3385.
[82] C.M. Pickart, R.E. Cohen, Nat. Rev. Mol. Cell Biol. 5 (2004) 177–187.
[83] K.Y. Sanbonmatsu, Curr. Opin. Struct. Biol. 22 (2012) 168–174.
[84] M. Mueller, S. Jenni, N. Ban, 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., Angew. Chem. Int. Ed. 45 (2006) 4064–4092.
[86] Y. Zhang, Curr. Opin. Struct. Biol. 18 (2008) 342–348.
[87] M. Christen, W.F. van Gunsteren, J. Comput. Chem. 29 (2008) 157–166.
[88] V. Tozzini, Curr. Opin. Struct. Biol. 15 (2005) 144–150.
[89] M. Müller, K. Katsov, M. Schick, J. Polym. Sci. B Polym. Phys. 41 (2003) 1441–1450.

[90] I. André, P. Bradley, C. Wang, D. Baker, Proc. Natl. Acad. Sci. U.S.A 104 (2007) 17656–17661.
[91] E. Mashiach-Farkash, R. Nussinov, H.J. Wolfson, Proteins Struct. Funct. Bioinf. 79 (2011) 2607–2623.
[92] A.M.J.J. Bonvin, Curr. Opin. Struct. Biol. 16 (2006) 194–200.
[93] S.E. Dobbins, V.I. Lesk, M.J.E. Sternberg, Proc. Natl. Acad. Sci. U.S.A 105 (2008) 10390–10395.
[94] M. Zacharias, Curr. Opin. Struct. Biol. 20 (2010) 180–186.
[95] E. Karaca, A.M.J.J. Bonvin, Structure 19 (2011) 555–565.
[96] S. Chaudhury, J.J. Gray, J. Mol. Biol. 381 (2008) 1068–1087.
[97] M. Król, R.A.G. Chaleil, A.L. Tournier, P.A. Bates, Proteins Struct. Funct. Bioinf. 69 (2007) 750–757.
[98] A.C. Steven, W. Baumeister, J. Struct. Biol. 163 (2008) 186–195.
[99] H.M. Berman, G.J. Kleywegt, H. Nakamura, J.L. Markley, Structure 20 (2012) 391–396.
[100] M.F. Lensink, R. Méndez, S.J. Wodak, Proteins Struct. Funct. Bioinf. 69 (2007) 704–718.
[101] C. Pons, M. D'Abramo, D.I. Svergun, M. Orozco, P. Bernadó, J. Fernández-Recio, J. Mol. Biol. 403 (2010) 217–230.
[102] D. Schneidman-Duhovny, M. Hammel, A. Sali, 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, Proteins Struct. Funct. Bioinf. 78 (2010) 3205–3211.
[105] A. Ahmed, P.C. Whitford, K.Y. Sanbonmatsu, F. Tama, J. Struct. Biol. 177 (2012) 561–570.
[106] M. Delarue, P. Dumas, Proc. Natl. Acad. Sci. U.S.A 101 (2004) 6957–6962.
[107] K. Hinsen, N. Reuter, J. Navaza, D.L. Stokes, J.-J. Lacapère, Biophys. J. 88 (2005) 818–827.
[108] K. Suhre, J. Navaza, Y.H. Sanejouand, Acta Crystallogr. Sect. D Biol. Crystallogr. 62 (2006) 1098–1100.
[109] K.-Y. Chan, L.G. Trabuco, E. Schreiner, K. Schulten, Biopolymers 97 (2012) 678–686.
[110] A.H. Ratje, J. Loerke, A. Mikolajka, M. Brünner, P.W. Hildebrand, A.L. Starosta, et al., Nature 468 (2010) 713–716.
[111] R.K.-Z. Tan, B. Devkota, S.C. Harvey, J. Struct. Biol. 163 (2008) 163–174.
[112] F. Tama, O. Miyashita, C.L. Brooks, J. Struct. Biol. 147 (2004) 315–326.
[113] F. Tama, O. Miyashita, C.L. Brooks, J. Mol. Biol. 337 (2004) 985–999.
[114] P.C. Whitford, A. Ahmed, Y. Yu, S.P. Hennelly, F. Tama, C.M.T. Spahn, et al., Proc. Natl. Acad. Sci. U.S.A 108 (2011) 18943–18948.
[115] W. Wriggers, R.A. Milligan, J.A. McCammon, J. Struct. Biol. 125 (1999) 185–195.
[116] S.C. Flores, M.A. Sherman, C.M. Bruns, P. Eastman, R.B. Altman, IEEE/ACM Trans. Comput. Biol. Bioinf. 8 (2011) 1247–1257.
[117] S.C. Flores, R.B. Altman, RNA 16 (2010) 1769–1778.
[118] F. Alber, S. Dokudovskaya, L.M. Veenhoff, W. Zhang, J. Kipper, D. Devos, et al., Nature 450 (2007) 683–694.
[119] D.J. Taylor, B. Devkota, A.D. Huang, M. Topf, E. Narayanan, A. Sali, et al., Structure 17 (2009) 1591–1604.
[120] K. Lasker, F. Förster, S. Bohn, T. Walzthoeni, E. Villa, P. Unverdorben, et al., Proc. Natl. Acad. Sci. U.S.A 109 (2012) 1380–1387.
[121] M. Campos, O. Francetic, M. Nilges, J. Struct. Biol. 173 (2011) 436–444.
[122] A. Loquet, N.G. Sgourakis, R. Gupta, K. Giller, D. Riedel, C. Goosmann, et al., Nature 486 (2012) 276–279.
[123] R. Das, I. André, Y. Shen, Y. Wu, A. Lemak, S. Bansal, et al., Proc. Natl. Acad. Sci. U.S.A 106 (2009) 18978–18983.
[124] W. Rieping, M. Nilges, M. Habeck, Bioinformatics 24 (2008) 1104–1105.
[125] M. Habeck, J. Struct. Biol. 173 (2011) 541–548.
[126] M. Habeck, W. Rieping, M. Nilges, Proc. Natl. Acad. Sci. U.S.A 103 (2006) 1756–1761.
[127] M. Hirsch, B. Schölkopf, M. Habeck, J. Comput. Biol. 18 (2011) 335–346.
[128] S.J. de Vries, M. van Dijk, A.M.J.J. Bonvin, Nat. Protoc. 5 (2010) 883–897.
[129] J. Fernández-Recio, M. Totrov, R. Abagyan, J. Mol. Biol. 335 (2004) 843–865.
[130] H. Hwang, T. Vreven, J. Janin, Z. Weng, 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, PLoS One 5 (2010) e12430.
[132] J. Seebacher, P. Mallick, N. Zhang, J.S. Eddes, R. Aebersold, M.H. Gelb, J. Proteome Res. 5 (2006) 2270–2282.
[133] P. Aloy, R.B. Russell, Nat. Rev. Mol. Cell Biol. 7 (2006) 188–197.
[134] H. Schreuder, C. Tardif, S. Trump-Kallmeyer, A. Soffientini, E. Sarubbi, A. Akeson, et al., Nature 386 (1997) 194–200.
[135] T. Kortemme, L.A. Joachimiak, A.N. Bullock, A.D. Schuler, B.L. Stoddard, D. Baker, Nat. Struct. Mol. Biol. 11 (2004) 371–379.