On Explorative and Integrative Modeling of Biomolecular Complexes

ISBN 987-90-393-6449-9

Doctoral Thesis
On Explorative and Integrative Modeling of Biomolecular Complexes
Gydo van Zundert
Computational Structural Biology, NMR Spectroscopy Research
Bijvoet Center for Biomolecular Research, Faculty of Science
Utrecht University, the Netherlands

Typeset with ConTEXt
Printed in the Netherlands by Uitgeverij BOXPress

On Explorative and Integrative Modeling of Biomolecular Complexes

Over Explorerend en Integratief Modelleren van Biomoleculaire Complexen

(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 25 november 2015 des middags te 4.15 uur

door

Gydo Cornelis Petrus van Zundert

geboren op 12 juni 1987
te Oud Gastel

Promotor:
Prof. dr. A.M.J.J. Bonvin

Dedicated to my parents,
the primum movens of my life.

Table of contents

List of abbreviations

ADH	Adipic acid dihydrazide
AIC	Average interactions per complex
AIR	Ambiguous interaction restraint
BS3	Bissulfosuccinimidyl suberate
CPU	Central processing unit
Cryo-EM	Cryo-electron microsopy
Cryo-ET	Cryo-electron tomography
CSP	Chemical shift perturbation
CW	Core-weighted
CXMS	Chemical cross-linking coupled with MS
DNA	Deoxyribonucleic acid
EPR	Electron paramagnetic resonance
FFT	Fast Fourier transform
FRET	Förster resonance energy transfer
GPU	Graphics processing unit
L	Laplace
LCC	Local cross-correlation
MS	Mass spectrometry
*N*D	*N*-dimensional
NMR	Nuclear magnetic resonance
PPDB	Protein-protein docking benchmark
RDC	Residual dipolar coupling
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
SAXS	Small-angle X-ray scattering
ZL	Zero-length

"Wie komm ich am besten den Berg hinan?"
Steig nur hinauf und denk nicht dran!

– Friedrich Nietzsche,
Die fröhliche Wissenschaft

Chapter 1

Introduction

1.1 Structural biology in the Omics-age

Since the start of modern-day Western science, Man is on a mission to thoroughly study Nature in order to understand, manipulate, and overcome her [1]. Above all, a fundamental insight into life is a hallmark in the whole scientific enterprise, where life is biologically represented in its irreducible form by the cell. The cell is a highly complex system that is regarded as the building block of life and is able to reproduce itself independently. Even though DNA holds a full blueprint of an organism, studied by the field of genomics, it is mainly the proteins that orchestrate the organization and functioning of cells, which has given rise to the field of proteomics, and the field of interactomics to characterize their interactions [2]. Recent technological and methodological advances have enabled the inquiry of the interaction networks that are formed by proteins, and showed that the set of all interacting protein complexes, the interactome, is 1 to 2 orders of magnitude larger than the total number of proteins that the genome encodes for, the proteome [3]. Inhibitors of these protein-protein interactions are an upcoming class of molecules with a profound impact on drug-development [4].

The field of structural biology tries to understand the workings of the molecules of life by studying their structure, preferable up to atomic resolution, as this provides a functional and mechanical description of the system [5] and a basis for rational drug design [6, 7]. Three dimensional, atomic-resolution structural information can be obtained by

high-resolution methods, mainly X-ray crystallography and NMR spectroscopy. Unfortunately, both methods are hampered by several limitations. X-ray crystallography is mainly limited by the production of high-quality crystals, an undertaking that becomes more difficult with increasing structure size, flexibility of the macromolecules, and transient complexes; for NMR spectroscopy it is mainly the size of proteins that is limiting structure determination, as spectra become heavily congested for larger complexes, making peak assignment infeasible. Furthermore, neither method is amenable to high-throughput investigations of complexes and large assemblies, a necessary requirement for the structural elucidation of the interactome.

In order to close the structure knowledge gap, computational methods have been devised to aid in this quest. Homology modeling is a successful approach to predict the structure of a protein with high-sequence identity to another already known structure, and heavily extends the structural knowledge of the proteome [8]. Macromolecular docking is the field that occupies itself with predicting the structure of a complex starting from their individual components [9], and can be divided in two main approaches: template based docking, similar to homology modeling, and "free" docking. It has been shown recently that templates are available for most complexes of structurally characterized proteins [10]. However, this approach is only amenable to complexes for which co-crystallized templates are available [11]. The "free" docking approach can be further subdivided into ab initio docking and data-driven docking. The former solely uses shape matching and physico-chemical principles to predict the structure of complexes with a limited success rate [12]; the data-driven approach tries to increase the success rate by including additional information from biophysical and biochemical methods during the docking [13, 14]. Data-driven docking is also more popularly known as hybrid or integrative modeling of biomolecular complexes.

1.2 Integrative modeling of biomolecular complexes

Integrative modeling is a procedure in which data from diverse sources are combined to accurately predict a model of a biomolecular complex [15, 16]. The procedure can be abstracted in four stages [17]:

1. *Gathering information:* collect information in the form of experimental data, bioinformatics predictions, statistical inference, or just about anything that can be of use during the modeling.

2. *Model representation and evaluation:* the degrees of freedom of the model should be chosen, i.e. using an all-atom model or a more coarse-grained representation, depending on how much information the data provide. In addition, scoring functions for each data-type need to be determined to indicate consistency between the models and the data.

3. *Sampling and optimization:* the sampling and optimization protocols should be chosen depending on the degrees of freedom of the system. For a 6 dimensional system, corresponding to the relative placement of two three-dimensional rigid bodies, an exhaustive search can be performed, while for higher-dimensional systems Monte Carlo and simulated annealing approaches would be more efficient.

4. *Scoring and analysis:* the resulting models need to be scored, ranked and clustered based on their congruency with the data to ascertain model precision and accuracy.

In the remainder of this section we will mainly describe sources of data to use during the modeling, and describe software packages that are geared towards integrative modeling.

1.3 Sources of information

In addition to the high-resolution structural techniques, many other experimental methods have been devised to extract structural or low-resolution information. NMR spectroscopy is also capable of pinpointing interface residues through the use of chemical shift perturbations (CSPs) [18], and the relative orientation of subunits to each other by residual dipolar couplings (RDCs) [19], among several other methods [20]. Small angle X-ray scattering (SAXS) experiments result in a 1D scattering curve, from which a diverse set of parameters can be determined with structural interpretation, e.g. radius of gyration, and even complete (low-resolution) shapes [21–23]. Biochemical methods such as mutagenesis and radical footprinting provide information on the binding interface.

Bioinformatics prediction methods can also deliver this information by analyzing sequences and extract conserved interface residues through co-evolution [24]. Two other experimental approaches that provide shape data and distance restraints are cryo-electron microscopy (cryo-EM) and chemical cross-linking coupled with mass-spectrometry (CXMS), which we will discuss more in-depth in the following.

1.3.1 Cryo-electron microscopy

Cryo-EM is a set of various transmission electron-microscopy techniques, namely cryo-electron tomography (cryo-ET), electron crystallography, and single-particle cryo-EM, that all ultimately results in a three dimensional density of the sample [25]. In cryo-ET whole cell slices are studied by systematically tilting and imaging projections of the sample; electron crystallography is mainly aimed at investigating membrane proteins that can form two-dimensional crystals; single-particle cryo-EM is used to study individual macromolecular assemblages by imaging many projections of random orientations of the assembly.

However, all three approaches are limited by the same phenomenon: the prolonged irradiation of the specimen with electrons results in extensive damage, reminiscent of the impact of a nuclear bomb [26]. To diminish this effect, the sample is typically plunge-frozen in liquid ethane to instantly vitrify and fixate it, resulting in a near-native hydrated state. However, the allowed electron dose is still severely limited, resulting in very noisy projections, well below atomic resolution. Electron crystallography tries to improve on this by using the high-resolution electron diffraction pattern to attain atomic resolution. Cryo-ET can significantly increase the resolution of particular assemblages by subtomogram averaging: a process where similar particles are aligned and averaged, resulting in an increased signal-to-noise ratio. Singe-particle cryo-EM in turn images many particles on a grid, each with a random orientation. By aligning similarly oriented projections, class averages can be obtained with a highly improved signal-to-noise ratio. If enough class averages are available, the three-dimensional density can be reconstructed through several iterative approaches.

Thanks to recent dramatic advances in direct electron detectors and improved particle processing software, the resolution of cryo-EM has impressively increased and sky-rocketed the cryo-EM field from blob-ology

[27] to the rising star in structural biology (subtitle of the cryo-EM Gordon Research Conference 2014), to revolutionizing structural biology [28, 29]. Although electron diffraction resolution has remained the same at around 2Å [30], cryo-ET's subtomogram averaging now attains sub-nanometer resolution [31], and the single-particle cryo-EM resolution record for now stands at 2.2Å [32].

Still, despite all these advances, the resolution of cryo-EM densities are in most cases typically too low for ab initio structural modeling. The information content of cryo-EM data is highly dependent on the resolution, with individual domains becoming visible at 15Å, secondary structure elements at 10Å for helices and 7Å for β-strands, and the separation of beta-sheets and bulky side-chains at around 4Å [33]. Thus, for typical cryo-EM data of 7Å resolution and lower, additional data need to be incorporated in an integrative approach to attain an atomic model of the macromolecular assembly.

1.3.2 Chemical cross-linking coupled with mass spectrometry

A very different method from cryo-EM is chemical cross-linking coupled with mass spectrometry. Here, protein complexes are covalently linked with chemical cross-links to determine spatial proximity between components. A standard CXMS experiment consists of six stages [34]: 1) the cross-links are added to the (purified) sample after optimizing the reaction conditions; 2) the cross-linked proteins are isolated to reduce the number of false-positives; 3) the cross-linked proteins are subsequently digested using trypsine or other proteases; 4) the resulting peptides are enriched using physico-chemical methods, such as size exclusion, affinity, and strong cation exchange chromatography. The final two steps are 5) MS optimization for peptide detection and 6) data-processing to detect cross-linked residues.

Even though the procedure is straightforward, each step is marked by optimization and many parameters need to be chosen, such as which linker to use, and how to enrich the cross-linked peptides [35, 36]. However, the major bottleneck is the final data analysis as millions to billions of fragments can be produced and need to be considered [34]. After a successful analysis, the cross-linked peptides can be mapped back on the proteins and distance restraints between components can be derived,

where the length and flexibility of the linker are used to define an acceptable range for the distance restraint. The shorter the linker the more information the restraint provides, though at the price of a reduced number of formed cross-links. So again, the inclusion of the low-resolution long-range distance restraints provided by CXMS require an integrative approach to accurately and precisely model the protein assemblies. A few recent examples where CXMS data were used are the INO80 complex of Saccharomyces cerevisiae [37], the Polycomb Repressive Complex 2 [38], and the 30S-elF1-elF3 translation initiation complex [39].

1.3.3 Software packages and platforms

Performing integrative modeling requires dedicated high-end software packages with powerful minimization, optimization and sampling algorithms. Currently, there are several software packages and platforms available that can handle data from a substantial number of experimental methods, but I will focus on three. One is the Rosetta software from the Baker lab [40], the second is the Integrative Modeling Platform (IMP) developed by the Sali lab [41], and the third is our in-house data-driven docking software HADDOCK (High Ambiguity Driven DOCKing) [42, 43].

Rosetta

Rosetta is at its core a structure prediction software package, and is well known for its elaborate and accurate scoring function and conformational sampling techniques [44]. Although originally a de novo protein prediction program [45], it has ventured into a more integrative approach and can now also perform X-ray crystallography refinement (MR-Rosetta) [46], use NMR data (CS-Rosetta), CXMS data [47], and recently also cryo-EM data [48, 49], resulting in the current prediction software package juggernaut that it is today [50]. The Rosetta source code was recently rewritten with the release of Rosetta3 [40]. Rosetta is free to use for academic purposes.

IMP

The IMP software package was from the outset designed as an integrative modeling platform, and is well-known for its use in the development of a model of the Nuclear Pore Complex [15, 51]. The IMP software consists of several user interface layers, each giving more control to the user [41, 52]. The base layer is written in C++ for speed, where each class is encapsulated for use in Python. This provides a scripting interface to setup an integrative modeling approach with data derived from diverse sources translated to restraints. One level higher are the direct user applications, such as MultiFit for cryo-EM [53, 54] and FoXS for the calculation of SAXS curves [55]. In addition, the IMP package is also integrated into the molecular graphics visualization program UCSF Chimera [56]. IMP is Free Software, licensed under the LGPL and GPL.

HADDOCK

The first version of HADDOCK was created in 2003, starting out as a binary protein docking program originally capable of incorporating CSP data and bioinformatics predictions [42]. Since then, HADDOCK's capabilities have steadily increased, and now also supports the use of RDCs [57], relaxation anisotropy [58], protein-DNA docking [59], solvated docking using explicit water [60, 61], docking up to 6 components [62], NMR pseudocontact shifts [63], SAXS and collision cross sections derived from MS [64], and protein-peptide docking [65]. The HADDOCK web server was introduced in 2010 [66] to provide a user-friendly interface to the science community. The HADDOCK software is free to use for academic purposes and ships with its source code, but does require CNS (Crystallography and NMR System) for its computational back end [67].

1.4 Explorative modeling

The goal of integrative modeling ultimately is to produce representative models of biomolecular assemblies that are consistent with the acquired data, thus putting the emphasis on the structural models. However, this does not necessarily provide insight into the information content of the restraints and certainty in the models. We can also turn this around, and instead put the emphasis on the data and aim at quantifying

the information content by counting all accessible states that are either consistent or inconsistent with the data. I am referring to this different paradigm and associated field as *explorative modeling*. A hallmark of this approach is to systematically sample a decent representative portion of the degrees of freedom of the system under investigation, and calculating for each sampled point the fit with the data, ultimately resulting in a distribution of states satisfying the input data. The method is inherently computationally demanding as the number of points to sample is sizable by itself and increases exponentially with the number of degrees of system being investigated. However, for two-body systems, corresponding to 6 degrees of freedom assuming rigid entities, the approach is manageable. The goal of explorative modeling is thus to provide the information content of the data, and preferably visualize this to the structural biologist, to aid in appreciating the impact of the data in restraining the accessible conformational/interaction space, to give insight into model uncertainty, and guide future work.

1.5 Overview of thesis

This thesis primarily describes new computational methods to handle cryo-EM and distance restraints data for integrative and explorative modeling. In **Chapter 2** I introduce a high-performance cross-correlation based rigid-body fitting software package called PowerFit to automatically fit high-resolution structures in low-resolution cryo-EM density maps. In addition to algorithm optimizations, it provides a novel and more sensitive scoring function to further extend the applicable resolution range. In **Chapter 3** I explore the resolution limits of rigid-body fitting in cryo-EM data and leverage this information to heavily accelerate the procedure through the use of multi-scale image pyramids. **Chapter 4** describes the incorporation of cryo-EM data in the HADDOCK software. The approach can be fully combined with all other available sources of information in HADDOCK, resulting in a truly integrative approach. Next, in **Chapter 5** I present the HADDOCK2.2 web server, an upgrade of the HADDOCK web server, for user-friendly integrative modeling of biomolecular complexes. **Chapter 6** deals with quantifying and visualizing the information content of distance restraints in general, and CXMS data in particular. It introduces the concept of the accessible interaction space, the set of all possible solutions of a complex, and

defines a way to exhaustively enumerate the accessible space. This is implemented in another software package called DisVis, and represents a first step into explorative modeling. I extend the approach further in **Chapter 7**, where interface residues are inferred from the accessible space defined by the distance restraints. The inferred residues can subsequently be used in the HADDOCK software to complement the docking process. In the final **Chapter 8** , I present a summary of the thesis and provide a personal perspective on the field of integrative modeling, proposing further lines of research.

1

Chapter 2

Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit

2.1 Introduction

Determining the architecture of large macromolecular complexes is of considerable interest to understand their function and mechanisms. Classical high-resolution methods such as X-ray crystallography and NMR spectroscopy might, however, struggle in doing that for large complexes that might be too flexible to crystallize or too large for peak assignment because of spectral overlap in NMR. Cryo-electron microscopy (cryo-EM) is quickly becoming the method of choice to gain structural insight into the nature of such large macromolecular assemblies. Especially with recent advances in detector technology and improved software and algorithms, the resolution of cryo-EM density maps is steadily increasing, occasionally at the point where models can be build in the density ab initio [28]. Still, for the bulk of the determined structures the level of detail is too low to routinely allow this and additional information is required to build an atomic representation of the system [68].

Typically, cryo-EM data are complemented with known high-resolution three dimensional (3D) models determined either experimentally or via homology modeling. These represent the pieces of the density puzzle that should all be fitted together in the map. The first step in the high-resolution modeling process is placing the subunits as rigid entities at the correct position in the density. This is often done manually using graphics software, most notably UCSF Chimera using its fit-in-map function [69]. This is unfortunate as it is subjective and can lead to over-interpretation

of the density map, as there is no objective scoring function to give an indication of the goodness-of-fit. This is especially problematic if flexible fitting is applied afterwards, since for the refinement to make sense the subunit should be located in a local minimum, else it might drift away from its initial position during the process. To this purpose a plethora of automatic rigid body fitting software has been developed [70]. A major class among those is the cross-correlation based programs, which are often combined with a full-exhaustive six dimensional (6D) grid search of the three translational and three rotational degrees of freedom [71–78]. This leads to a thorough and objective analysis of all possible solutions to locate the global cross-correlation minimum.

The first full-exhaustive cross-correlation based software was published by Volkmann and Hanein [71]. The approach was further developed by Chacón and Wriggers [73] using the Fast Fourier Transform (FFT) algorithm in combination with the cross-correlation theorem, which decreases the computational complexity of the search. In addition, they applied a Laplace pre-filter on the density and search object, significantly extending the applicable resolution range. Roseman [72] introduced the more sensitive local cross-correlation (LCC) score to fit subunits instead of whole complexes in the density. Wu et al. [75] acknowledged the problem of overlapping densities of neighboring subunits at lower resolutions and developed a core-weighted (CW) cross-correlation score to minimize this effect by biasing the weight of density toward the core of the search object. Recently, Hoang et al. [78] implemented a GPU-hardware-accelerated version based on FFT techniques to calculate the LCC score, building on the earlier work by Roseman [79].

Here we report on further developments in cross-correlation based rigid body fitting. In the Methods section, we first shortly describe the essence of exhaustive cross-correlation based fitting and introduce a new cross-correlation function that combines the core-weighted approach of Wu et al. [75] with the LCC, demonstrating how it can be calculated using FFTs. Furthermore, to decrease the time required to perform a full exhaustive search we use the optimal rotation sets developed by Karney [80] and decrease the size of the density by automatically resampling the data, if possible, and trimming padded regions. In the Results section, we investigate the sensitivity of the newly developed scoring function by automatically fitting the subunits of the 80S D. melanogaster ribosome

[81]. Lastly, we present a performance comparison against other fitting software using the GroEL/GroES system with experimental data [82].

We implemented our approach in a Python software package called PowerFit, which can run on multi-core CPU machines and can be GPU-accelerated using the OpenCL framework. PowerFit has been tested on Linux, MacOSX and Windows operating systems and is Free Software. The source code with detailed installation instructions and application examples can be found at `https://github.com/haddocking/powerfit`.

2.2 Methods

2.2.1 State of the art of rigid body cross-correlation based fitting

The goal of cross-correlation rigid body fitting is to determine the three translational and three rotational degrees of freedom of the model that optimize the cross-correlation score between the high-resolution model and the density. To this end, the model is first blurred to the resolution of the cryo-EM data to properly calculate the goodness-of-fit. It should be noted that, although the notion of the exact resolution of a cryo-EM density is still a matter of debate and can actually be anisotropic, the reported resolution of the data is usually sufficient for fitting purposes. This blurred model is then fitted by performing a systematic, full-exhaustive search of the 6D space and saving locations corresponding to high cross-correlation values. Predictably, the problems with this approach are sensitivity of the scoring function and speed of the search. The sensitivity of the global cross-correlation score as originally used by Volkmann and Hanein [71] is often compromised as, typically, subunits instead of the whole complex are fitted into the density. To make things worse, at lower resolution the local densities of neighboring subunits are overlapping, resulting in systemic noise mainly at the edges of the search model. To overcome the first problem, Roseman [72] introduced the local cross-correlation function, which effectively is the cross-correlation normalized under the running footprint of the shape of the model. This localizes the score to only the region of interest, making the fitting of subunits feasible. As for the effect of overlapping densities of neighboring subunits, this can be minimized by biasing the density toward the core of the search object. Wu et al. [75] incorporated this concept by calculating

the core-index of each voxel of the search object, where the core-index is a measure for how far the voxel is from an edge. To further enhance the sensitivity of the scoring function, a Laplace pre-filter can be applied to the cryo-EM density and search object [73]. Originally combined with the global cross-correlation, it was recently shown that combining it with the LCC further extends the applicable resolution range [78].

To increase the efficiency of the search and minimize computational costs, the main innovation was the use of the cross-correlation theorem in combination with FFTs. By discretizing the model density on a grid with the same voxel spacing and size as the cryo-EM grid, a translational scan can be performed using the FFT-accelerated approach. This reduces the computational complexity from N^2 to $N \log(N)$, where N is the total number of voxels of the cryo-EM data. After each translational scan, the model density is rotated and the process is repeated until a pre-set rotational sampling density is achieved, meaning that the time required for a search depends linearly on the number of rotations sampled. The rotation step can be accelerated by directly rotating the density of the search object instead of repeatedly rotating the high-resolution model and blurring it afterwards. The GPU-architecture especially is suited for this task as tri-linear interpolation can be done with high-efficiency [78].

2.2.2 Increasing the sensitivity by combining the LCC with the core-weighted approach

Originally the core-weighted procedure was combined with the global cross-correlation, which significantly extended the resolution range in which a subunit could be successfully fitted into the density. The same procedure is expected to also improve the sensitivity of the better performing LCC. Combining the two approaches results in what we defined here as the core-weighted LCC (CW-LCC) scoring function:

$$\text{CW-LCC} = \frac{1}{N} \frac{\sum_i^N \left(w_i \rho_c - \overline{\rho_c^w}\right) \cdot \left(w_i \rho_o - \overline{\rho_o^w}\right)}{\sigma_c^w \sigma_o^w} \tag{2.1}$$

where the summation is over all the N voxels that are within a distance of half the resolution of any atom of the search object indexed by i; w_i is the core-index of voxel i; ρ_c and ρ_o are the intensities of the search object and the cryo-EM density at voxel$i$, respectively, $\overline{\rho_c^w}$ and $\overline{\rho_o^w}$ are

the core-weighted density average for the search object and the local cryo-EM density, respectively, given by $\overline{\rho_x^w} = 1/N \sum_i^N w_i \rho_x$. σ_c^w and σ_o^w correspond to the core-weighted density standard deviation given by $\sigma_x^w = \sqrt{\left(\overline{\rho_x^w}\right)^2 - \overline{\left(\rho_x^w\right)^2}}$ with $\overline{\left(\rho_x^w\right)^2} = 1/N \sum_i^N \left(w_i \rho_x\right)^2$. The CW-LCC reduces to the regular LCC by setting $w_i = 1$. The Laplace pre-filtered scoring function is defined by performing the mapping $\rho_x \rightarrow \nabla^2 \rho_x$ in **Eq. 2.1**. In order to calculate the CW-LCC we first need to define the core-index of each voxel.

Determining the core-index w_i

The core-index is a measure for how close a voxel is to the core of the density of the subunit that is being fitted. We calculate the core-index by progressively eroding a binary mask of the search object and summing each eroded mask together, see **Figure 2.1**A for a 2D example. This guarantees that voxels at the surface have a low core-index value, while voxels deeply buried get a higher value, even for complex shapes.

Using Fourier techniques to calculate the CW-LCC

Starting from **Eq. 2.1** and following in the spirit of Roseman [79], we can normalize the core-weighted density $w_i \rho_c$ of the template by setting $\overline{\rho_c^w} = 0$ and $\sigma_c^w = 1$, which simplifies **Eq. 2.1** to

$$\text{CW-LCC} = \frac{1}{N} \frac{\sum_i^N \rho_c^n \cdot w_i \rho_o}{\sqrt{\left(\overline{\rho_o^w}\right)^2 - \overline{\left(\rho_o^w\right)^2}}} \tag{2.2}$$

where ρ_c^n indicates the normalized core-weighted density. This leaves three terms to be determined: the nominator, which we refer to as the core-weighted global cross-correlation (CW-GCC); the square of the average core-weighted density, $\left(\overline{\rho_o^w}\right)^2$, and the average of the squared core-weighted density, $\overline{\left(\rho_o^w\right)^2}$, of the cryo-EM data. These can be calculated using FFTs as follows

$$\text{CW-GCC} = \mathcal{F}^{-1}\left[\mathcal{F}\left(w\rho_c^n\right)^* \times \mathcal{F}\left(\rho_o\right)\right] \tag{2.3}$$

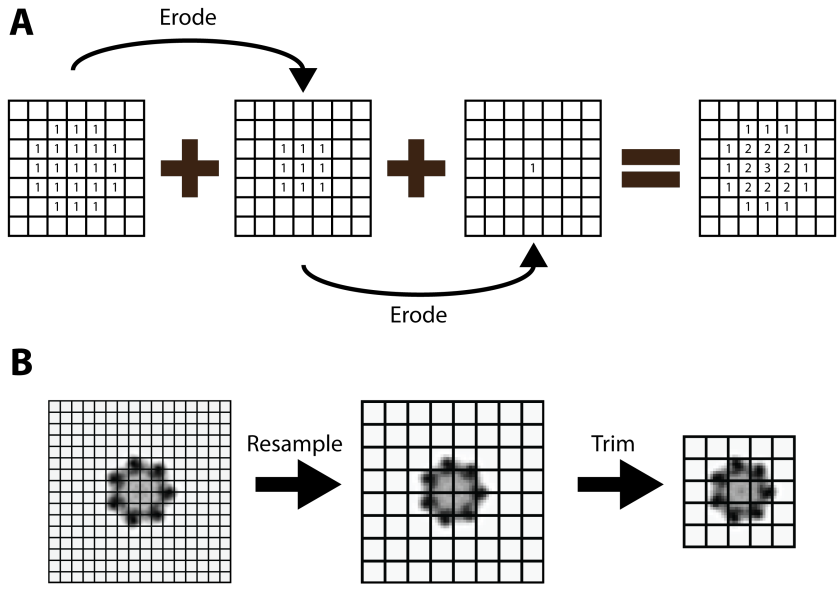

Figure 2.1 Algorithms used in PowerFit. (A) Illustration of the calculation of the core-weighted mask. The initial binary mask is progressively eroded and summed. (B) Illustration of the impact of resampling and trimming on a slice of the GroEL/GroES density where each square consist of 8×8 voxels. After resampling and trimming the final size is significantly reduced.

$$\left(\overline{\rho_o^w}\right)^2 = \mathcal{F}^{-1}\left[\mathcal{F}\left(w\right)^* \times \mathcal{F}\left(\rho_o\right)\right]^2 \tag{2.4}$$

$$\overline{\left(\rho_o^w\right)^2} = \mathcal{F}^{-1}\left[\mathcal{F}\left(w^2\right)^* \times \mathcal{F}\left(\rho_o^2\right)\right] \tag{2.5}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the Fast Fourier transform and its inverse, respectively, * is the complex conjugate operator, $\times$ is the element wise multiplication operator, w is the core-weighted mask, ρ_c and ρ_o are the calculated and experimental densities, respectively. In **Eq. 2.3** it is the search object that is multiplied with the core-weighted mask, instead of the cryo-EM density. It is this trick which allows the CW-GCC to be calculated using FFTs. Note that even though there are 9 Fourier transforms required to calculate the CW-LCC, only 6 need to be calculated for every orientation sampled, as the 3 Fourier transforms of the cryo-EM data can be calculated just once before the search. So the FFT-accelerated CW-LCC effectively costs only one Fourier transform more than the regular LCC [78].

2.2.3 Speeding up the search

Using optimized rotation sets to limit rotational degeneracy

Since the computational complexity of the exhaustive search depends linearly on the number of rotations sampled, optimizing and limiting rotational degeneracy is important for an efficient search. However, sampling rotations or orientations in a systematic and efficient manner is a non-trivial exercise. As such, the number of orientations that are sampled to guarantee a certain rotational sampling density can differ widely. For example, COLORES uses proportional Euler angles [73], while gEMFitter performs an icosahedral tessellation to generate rotations [78], resulting in 1264 and 900 orientations sampled for a coarse 24° search, and 119664 and 92160 for a fine 5° search, respectively.

In our implementation, we make use of the optimal rotation sets determined by Karney [80], originally developed for solid state NMR. These sets were pre-calculated by enclosing the hypersphere of unit quaternions and require only 648 orientations for a 20.83° search and 70728 orientations at a 4.71° sampling rate. This is an enhancement of the sampling efficiency of at least a factor of 1.3 compared to gEMFitter, while offering a denser rotational sampling interval.

Decreasing the map size by resampling and trimming the density

In addition to the number of rotations sampled, the computational complexity of the search scales with $N \log(N)$ where N is the number of voxels of the data. This is the major determinant for the computational resources required. Limiting the density size is thus key to limiting the time required for a search. Cryo-EM data are often oversampled with respect to their resolution incurring a significant computational cost to perform an exhaustive search. Because neighboring voxel intensities will be highly correlated, resampling the cryo-EM data will not affect the scoring sensitivity significantly. However, as there is still signal after the resolution cutoff, resampling the cryo-EM data to Nyquist rate will introduce aliasing effects and image distortions. Therefore, we choose to resample the cryo-EM map to a default rate of 2 times Nyquist, i.e. the data are resampled such that the voxel spacing is 1/4th of the resolution, allowing for a safe buffer to minimize aliasing effects.

In addition to that, cryo-EM data are usually generously padded with voxels containing only noise. It is not uncommon for the padding to increase the number of voxels in each direction by a factor of 2 or more. This comes at a considerable cost when performing an exhaustive search as the number of voxels grows by a factor of 8 or more. To eliminate the computational cost incurred by this padding, we trim the padded voxels. The effect of resampling and trimming is shown in **Figure 2.1**B on a slice of the GroEL/GroES complex (EMD-1046).

2.2.4 Implementation and availability

We implemented our methods in a Python package named PowerFit that comes with a command line tool eponymously named *powerfit*. A flowchart of the *powerfit* algorithm is shown in **Figure 2.2**. It requires as input a PDB structure, a cryo-EM map and its resolution. Optional parameters are the rotational sampling density (default=10.83°), whether to resample and/or trim the density and use the Laplace pre-filter and/or core-weighted procedure, and the number of PDBs that should be written to file after the search. In addition, the number of CPU processors available to the search can be specified or whether the computations should be off-loaded to the GPU.
After invoking *powerfit*, the software will first try to resample the cryo-EM map to 2 times Nyquist rate and then trim it. A density of the search object (the 3D structure) is constructed by a Gaussian convolution where the standard deviation is a function of the resolution. Also, a binary mask is computed out the structure, where voxels within half a resolution distance from any atom in the model are set to 1 and otherwise 0. Both the search object density and mask are discretized on grids of equal sizes and spacing as the cryo-EM density map to allow for an FFT-accelerated search. The Laplace pre-filter is applied on the cryo-EM and template densities, if requested. A core-weighted mask is calculated from the initial binary mask using the procedure described above. The data necessary for the search are offloaded to the GPU if requested. The template and mask are rotated using tri-linear interpolation, where texture memory acceleration as described by Hoang et al. [78] is used when possible. For each rotation sampled, a translational correlation scan is performed using FFTs. The rotational solution with the highest score is saved at every grid position. This continues until the

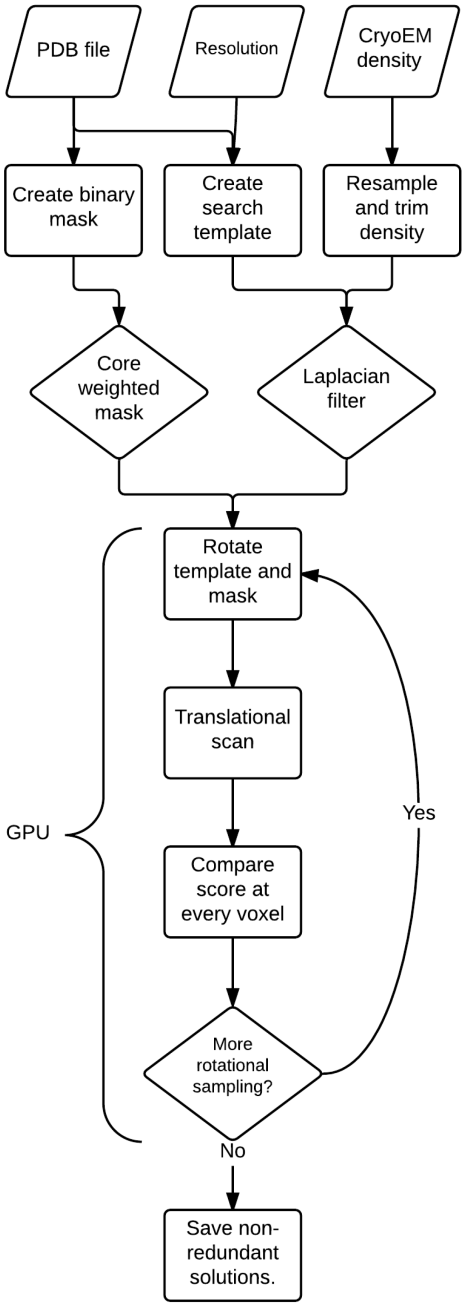

Figure 2.2 Flowchart of the powerfit algorithm.

requested rotational sampling density is achieved. At the end, the grid, which contains at each position the highest found cross correlation score for all sampled rotations, is segmented using a 3D watershed algorithm [83] in order to remove redundant solutions. The location of each maximum together with its correlation score and corresponding rotation are written to file as well as the corresponding PDB coordinates of the top N solutions (where N is a user-defined parameter).

PowerFit is written in the Python language (Python2.7) and requires the NumPy and SciPy packages. The CPU version can be further accelerated by installing the FFTW3.3 library together with pyFFTW. To offload the computationally intensive search to the GPU, we used the OpenCL framework together with the clFFT library, a high-performance FFT library for OpenCL. Python bindings were available through the pyopencl and gpyfft packages. PowerFit is licensed under the MIT license and can be downloaded from `https://github.com/haddocking/powerfit` together with instructions on how to install and use it. It has been successfully tested on Linux, MacOSX and Windows systems and its GPU-accelerated version can run on both AMD and NVIDIA GPUs, minimizing vendor lock-in.

2.3 Results

2.3.1 Scoring sensitivity of the core-weighted LCC

To test the scoring sensitivity of the CW-LCC, we used PowerFit to fit each subunit of the 80S D. melanogaster ribosome [81] independently in the density at different resolutions. To this end, we simulated cryo-EM data from a deposited model (4V6W) from 6 to 30Å resolution in 1Å increments. The cryo-EM data were created using a Python script based on the molmap function in UCSF Chimera [69]. Subsequently, we fitted each subunit using the LCC and CW-LCC score and also together with the Laplace pre-filter (L-LCC and L-CW-LCC) resulting in four different scoring functions. As there are 86 subunits in the assembly, we performed 8600 exhaustive searches in total (86 subunits $\times$ 25 resolutions $\times$ 4 different scores). The voxel spacing of the simulated data was 1/4th of the resolution with a maximum of 4Å using a rotational sampling density of 20.83° (648 rotations). We defined a fit as successful if the positional

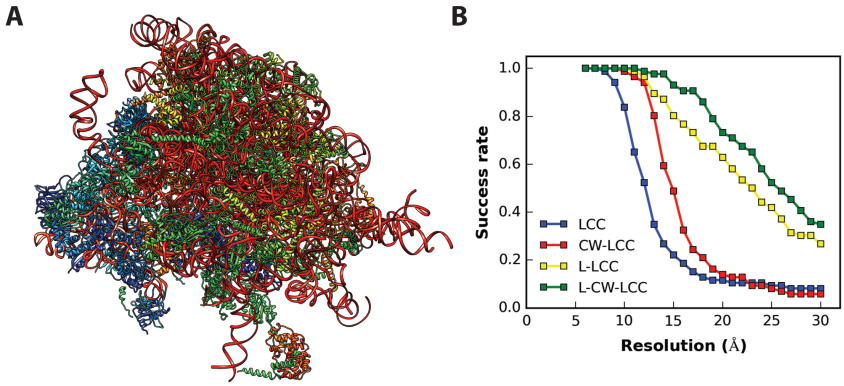

Figure 2.3 Fitting subunits of the ribosome in simulated cryo-EM data. (A) The 80S ribosome assembly of D. melanogaster (4V6W). (B) The success rate from the fitting of 86 individual subunits is plotted versus the resolution of the cryo-EM data for the four different scoring functions (LCC = local cross correlation; CW-LCC = core-weighted LCC; L-LCC; Laplace pre-filtered LCC; L-CW-LCC = Laplace pre-filtered CW-LCC).

RMSD of a solution in the top 10 was smaller than 8Å compared to the reference structure (4V6W), which is a reasonable 2 voxel spacings away from the correct solution at 16Å resolution and lower. Since we were testing the sensitivity of the scoring function, the orientation of the correct model was included in each search. The results of the scoring comparison are shown in **Figure 2.3**B.

All four scoring functions can fit all subunits correctly in the density at 6Å and 7Å resolution. However, the performance of the LCC begins to decrease after 8Å resolution and the number of successful cases drops markedly up to 18Å resolution, to further only decrease. The CW-LCC score performs significantly better, only starting to drop at 10Å resolution. After that, it follows a similar pattern as the LCC with a quick drop first and a more stable region in the end. The core-weighted approach extends the applicable resolution range of the LCC by a respectable 3Å. The scoring functions combined with the Laplace pre-filter are evidently performing better. The L-LCC score is almost 100% successful up to 12Å resolution. The success rate drops at lower resolutions, though not as fast as the LCC and CW-LCC score and follows a rather linear trend, which is in contrast with the other scoring methods. The best performing score is the L-CW-LCC as expected. It is capable of fitting all subunits

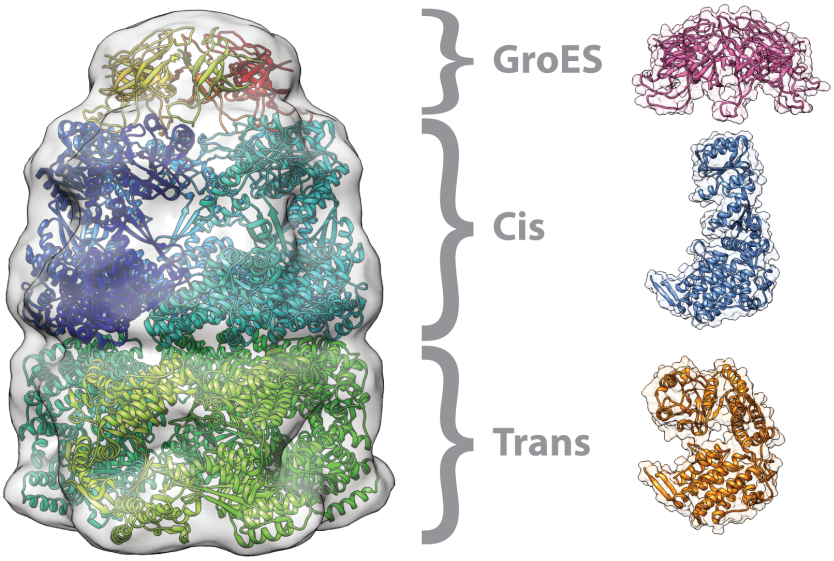

Figure 2.4 The GroEL/GroES density (EMD-1046) with its reference structure fitted inside (1GRU). The subunits used in the full exhaustive search are shown on the right.

up to a resolution of 12Å and is near-perfect up to 15Å resolution. Similar to the L-LCC score, the success rate drops linearly up to 30Å resolution.

This analysis demonstrates that including both the Laplace pre-filter and the core-weighted approach results in the most sensitive scoring function. The Laplace pre-filter seems to have the largest impact, changing the drop rate of the curve to a linear one, while the inclusion of the core-weighted approach results in a right shift of the curve.

2.3.2 Fitting performance of *powerfit*

Fitting subunits in the GroEL/GroES complex

As an experimental test case for *powerfit*, we used the GroEL/GroES complex (EMD-1046, **Figure 2.4**) [82], which has been used in the cryo-EM modeling challenge and makes comparison with other software possible [78, 84]. The crystal structure of GroEL/GroES (1GRU) was used as a reference. We fitted a subunit of the trans and cis rings of GroEL and the whole GroES ring (as with other software attempts, fitting individual subunits of GroES was not successful [78]) independently in the

Table 2.1 Fitting performance on the GroEL/GroES complex of the Laplace pre-filtered (L) and core-weighted local cross-correlation (CW) score. The average RMSD of fitted subunits are shown in Å when using the full and resampled map.

	Trans		Cis		GroES	
	L	CW	L	CW	L	CW
Full map	2.9	3.4	4.6	3.8	4.6	4.2
Resampled	5.5	5.2	7.6	7.3	4.4	4.4

density, using the four different scoring functions, with and without resampling. The rotational sampling density was set at 4.71°. For the cis and trans rings we took the top 7 best scoring fits and calculated the average RMSD to the 1GRU reference structure; for the GroES ring we took the best fit only. The results are shown in **Table 2.1**.

The LCC score was incapable of fitting any subunit properly as was noted earlier [78]. In case of the GroES lid, it actually places it upside-down in the density. The CW-LCC is more successful in this respect, and properly fits the GroES ring at the top of the density with an RMSD of 7.4Å using the full map and 4.4Å when using the resampled map. However, it still fails to accurately fit the trans and cis subunits in the density. In general, the Laplace pre-filter scoring functions are capable of fitting all subunits successfully in the density, with no significant difference in accuracy considering the resolution of the data. As expected, the accuracy lowers when we resample the map to two times Nyquist, though the difference is less than one voxel spacing; when refitting the top 7 solutions using one translational scan in the fitted orientation with the regular voxel spacing, similar results are obtained, but at a markedly lower computational cost (see next section). The fitting results from *powerfit* (RMSD of 3.4, 3.8 and 4.2Å) are competitive compared to previous published ones: gEMfitter reported an RMSD of 2.8, 4.0 and 5.3Å for the trans, cis and GroES ring [78], respectively, and Segger 3.1, 5.1 and 6.0Å [85].

Timing comparison of powerfit

We also investigated the effect of trimming and resampling the density on the time required to perform a run. As the Laplace pre-filter only

Table 2.2 Time required for a coarse (20.81°) and fine (4.71°) rotational search on the GroEL/GroES complex using the local cross-correlation score. The timings of fitting in the full, trimmed, and resampled and trimmed (R + T) map version are reported.

	Coarse		Fine	
	CPU	GPU	CPU	GPU
Full map	3m 32s	18s	6h 23m	23m 50s
Trimmed	58s	7s	1h 37m	3m 54s
R + T	10s	4s	13m 6s	1m 6s

needs to be applied once, the timings of the regular and Laplacian scores are similar. We therefore only show times for the L-LCC and L-CW-LCC scores. The results of the timing runs are shown in **Table 2.2**.

Running a coarse 20.81° rotational search can be done in a few minutes, even on a single processor with a map size of 128 × 128 × 128 voxels. However, for a fine rotational sampling density of 4.71° an exhaustive search already requires more than 6 hours. Using a GPU (NVIDIA Geforce GTX 680) to accelerate the search reduces the time required to approximately 30 minutes. When trimming the density before the search, which in the GroEL/GroES case reduces the map size to 72 × 72 × 90, the time required for a fine search drops to ~1.5 to 2 hours on a single processor and only 5 minutes on a GPU. It should be emphasized that trimming the map does not have any impact on the search accuracy and thus should always be applied for a faster search. Further minimizing the map size by resampling the density results in 36 × 36 × 45 voxels, and only requires 15 minutes on a single CPU and 1 to 2 minutes on a GPU. Thus, we advise to always use the trimming option and start a search using the resampled option. The resulting solutions can than be refitted using a single translational scan on the non-resampled map for an optimal speed to accuracy trade-off.

We compared the fitting times of *powerfit* against another GPU-accelerated rigid body fitting software gEMfitter [78]. The results are shown in **Table 2.3**. Running gEMfitter using a 5° sampling density (92160 orientations) with the L-LCC scoring functions, requires 5 hours and 48 minutes against 6 hours and 23 minutes for *powerfit*, without any of the simplifications introduced here, on a single processor (Intel Core i7-3632QM). As the bulk of the time is spent on computing FFTs, the

Table 2.3 Timing comparison between *powerfit* and gEMfitter using a fine rotational search on the GroEL/GroES complex. Both softwares offer the user the option to resample the density.

	CPU		GPU	
	gEMfitter	*powerfit*	gEMfitter	*powerfit*
Full map	5h 48m	6h 23m	11m	25m 48s
Resampled	38m 2s	41m 25s	-	1m 40s

difference in performance might be found in the fact that the gEMfitter binary has been compiled with the mkl-library and *powerfit* with GCC. gEMfitter also has a resampling option, which reduces the running time to 38 minutes. Only applying the resampling option reduces the running time for *powerfit* to 41 minutes, and combined with trimming the running time drops further to 13 minutes using the same L-LCC scoring function. We could not properly compare the GPU-accelerated version of gEMfitter against *powerfit* as the provided gEMfitter binary runs only on Ubuntu systems with NVIDIA GPUs and was not at the authors' disposal. However, the gEMfitter article reports 11 minutes running time using a NVIDIA C2075 GPU, which is significantly shorter than *powerfit* without trimming and resampling. With the latter two options turned on, the *powerfit* timings drop to close to 1 minute on a GTX 680 GPU card. Again, since the bulk of the time is spent on computing FFTs, the difference probably arises in the efficiency of the FFT implementation: the CUDA FFT implementation is specifically optimized for NVIDIA GPUs while the clFFT implementation is mainly optimized for AMD architecture, but runs on all OpenCL supported architectures. So there is a choice between performance versus portability, although, with trimming and resampling enabled, *powerfit* is still faster.

Additional complexes fitted with powerfit

To validate our approach further, we applied *powerfit* on three additional cases in the resolution range of 8.9 to 13.5Å (**Table 2.4**, **Figure 2.5**). EMDB entry 2325 is another GroEL/GroES complex, but at a considerably higher resolution of 8.9Å compared to the 1046 density [86]. The increased level of detail allowed to fit each GroES subunit independently in the map, irrespective of the scoring function used, with the

Table 2.4 Additional complexes fitted with *powerfit*. A fine rotational search (4.71°) was performed using the four scoring functions (LCC = local cross correlation; CW-LCC = core-weighted LCC; L-LCC = Laplace pre-filtered LCC; L-CW-LCC = Laplace pre-filtered CW-LCC).

EMDB	Resolution (Å)	PDB	Score	RMSD (Å)	Rank
2325	8.9	3ZPZ:O	LCC	1.7	1 – 7
			CW-LCC	1.9	1 – 7
			L-LCC	1.7	1 – 7
			L-CW-LCC	1.5	1 – 7
1884	9.8	2YKR:W	LCC	2.5	1
			CW-LCC	3.0	2
			L-LCC	2.2	1
			L-CW-LCC	2.2	1
2017	13.5	4ADV:V	LCC	60.8	1
			CW-LCC	1.3	1
			L-LCC	5.9	1
			L-CW-LCC	4.7	3

correct 7 fits found in the top 7. The other two cases are ribosomes with a GTPase [87] and methyltransferase [88] bound to it, subunits with comparable size. For entry 1884 with a reported resolution of 9.8Å, the RsgA GTPase was correctly fitted in the density by all four scoring functions and was found within the top 2 best scoring solutions. The ribosome map 2017 with the bound KsgA methyltransferase has a somewhat lower resolution of 13.5Å. In this case, the LCC was incapable of correctly fitting the subunit in the density. The other scoring functions placed the subunit properly in the map, with the correct fit found within the top 3 best scoring solutions.

2.4 Conclusion

In this work we have introduced PowerFit, an open source Python package, which comes with a command line tool *powerfit* to perform an exhaustive cross-correlation based rigid body search. It implements a

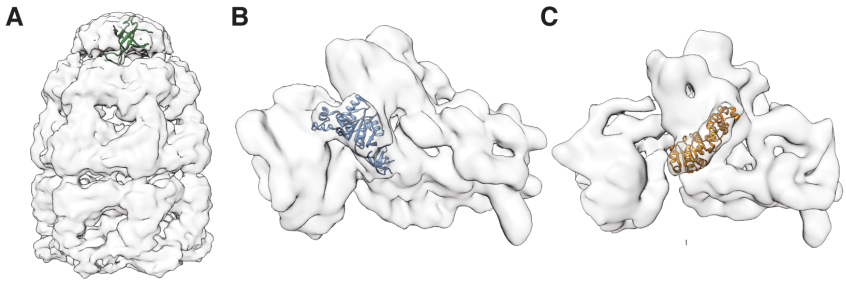

Figure 2.5 Cryo-EM densities together with the subunits that were independently fitted. (A) GroES subunit in GroEL/GroES complex (EMD-2325), (B) RsgA GTPase in 30S ribosome (EMD-1884), and (C) KsgA methyltransferase in 30S ribosome (EMD-2017).

new core-weighted enhanced LCC score that significantly expands the applicable fitting resolution range. In addition, *powerfit* minimizes the computational time requirements by using optimized rotation/orientation sets, trimming and resampling the electron density, and leveraging the computational resources provided by GPUs. PowerFit is therefore a valuable addition to the structural biologist toolbox, allowing obtaining an objective initial fit of high-resolution subunits in low-resolution cryo-EM density maps within a reasonable time.

Notes

This Chapter is based on: G.C.P. van Zundert and A.M.J.J. Bonvin. Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. *AIMS Biophysics* **2**, 73–87 (2015). Furthermore, dr. Marcus Weingarth is acknowledged for making GPU resources available.

2

Chapter 3

Leveraging the limits of rigid-body fitting in cryo-EM maps using multi-scale image-pyramids

3.1 Introduction

A structural understanding of large macromolecular complexes is of fundamental importance to rationalize and manipulate cellular processes. Cryo-electron microscopy (cryo-EM) is quickly becoming the method of choice for studying these macromolecular machines as recent advances are enabling unprecedented levels of detail to be visualized [28]. Sub-nanometer resolution maps are no exception anymore, although the level of detail is usually still too low for de novo building of atomic structures. When possible, cryo-EM data are therefore combined with high-resolution atomic models of subunits for a proper structural understanding of the data. Typically, the first step in the modeling process is placing the subunits in the density as rigid bodies, after which the models can be refined using some flexible fitting procedure [70].

A variety of tools and software have been developed to help users in the rigid body fitting, both for manual and automatic placement. Though manual placement is frequently performed, most notably using UCSF Chimera [69], it is subjective and can lead to over-interpretation of the data, as there is no objective target function to be optimized. The available local cross-correlation function in UCSF Chimera is limited, as it samples only the current orientation. The problem of manual fitting is exacerbated when flexible fitting is applied afterwards, as it requires

an initial local cross-correlation minimum between the model and the density, else the model would drift away from its fitted location.

An automatic and objective method to determine the placement of the subunits is to perform a full-exhaustive systematic cross-correlation search of the three translational and three rotational degrees of freedom of the model in the density. Many advances have been made in both sensitivity and speed of cross-correlation based rigid body fitting [71–73, 75–78, 89, 90]. In **Chapter 2** we introduced the core-weighted local cross-correlation scores in our rigid-body fitting package PowerFit. However, to our knowledge, no thorough investigation into the limits of rigid body fitting has been performed so far, nor has the resolution requirements to fit a subunit of a certain size in the density been quantified. In addition, as the size of cryo-EM data has been steadily increasing as a result of the higher information content, the CPU requirements for an exhaustive search, which is usually performed using Fast Fourier Transform (FFT)-techniques for fast translational scans, are considerably increasing, which slows down the entire process.

Here we report on a comprehensive exploration of cross-correlation based rigid-body fitting into cryo-EM densities, using five high-resolution ribosome maps in the range of 5.5 to 6.9Å for which high-resolution models are available. We analyze the success rate of fitting all 379 subunits into these maps as a function of resolution using four different scoring functions. This is done by progressively lowering the resolution of the initial data down to 30Å. Furthermore, we show how the size of the subunits influences the success rate of fitting and how over-interpreted regions of the map can be identified. Finally, we leverage this information by using the concept of multi-scale image pyramids [91], well known in the field of image analysis, to significantly reduce the required computational resources and time to perform a fit by up to two orders of magnitude. This is implemented in our PowerFit package for fast rigid body fitting in cryo-EM data, which can be freely downloaded from `https://github.com/haddocking/powerfit`.

Table 3.1 The five high-resolution ribosome cases with deposited atomic models.

EMDB-ID	Resolution (Å)	PDB-ID	Number of subunits
1780 [94]	5.5	4V7E	88
2620 [95]	6.9	4UJE	83
2845 [96]	6.5	4UER	39
5591 [81]	6.0	4V6W	86
5976 [97]	6.2	3J77	83

3.2 Methods

3.2.1 Exploring the limits of rigid-body fitting

To explore the resolution limit for successful rigid body fitting, we selected five high-resolution cryo-EM ribosome maps from the EMDatabank [92], ranging in resolution from 5.5 to 6.9Å, for which structural models were deposited in the Protein Databank [93] (**Table 3.1**). The ribosome is an excellent case study as it contains many chains of various sizes and types. Subsequently, we tried to fit each separate chain independently in their respective density map with PowerFit, using four different scoring functions: the local cross-correlation (LCC), the core-weighted (CW-) LCC, and their Laplace pre-filtered versions the L-LCC and L-CW-LCC, respectively, given by the master equation

$$\mathrm{CC} = \frac{1}{N} \frac{\sum_i^N (w_i \rho_c - \overline{\rho_c}) \cdot (w_i \rho_o - \overline{\rho_o})}{\sigma_c^w \sigma_o^w} \tag{3.1}$$

where the summation is over all N voxels that are within half a resolution distance of any atom of the search object indexed by i; w_i is a weight factor given to voxel i; ρ_c and ρ_o are the intensities of the search object and the cryo-EM density at voxel i, respectively; $\overline{\rho_c^w}$ and $\overline{\rho_o^w}$ are the weighted density average for the search object and the local cryo-EM data, respectively, given by $\overline{\rho_x^w} = 1/N \sum_i^N w_i \rho_x$. Finally, σ_c^w and σ_o^w are the weighted density standard deviations for the search object and EM-data. The LCC-score is defined by setting w_i to 1, while for the CW-LCC w_i is given by the core-index (Wu et al. [75], **Chapter 2**), a

measure for how close the voxel is to the core of the search object. The Laplacian enhanced scoring functions are defined by mapping $\rho \rightarrow \nabla^2 \rho_x$ [73].

After each round of fitting, the resolution of the cryo-EM data was lowered by 1Å using the following procedure. Assuming that the density is described by a collection of atoms with a spherical Gaussian density distribution, where the width of the Gaussian depends on the resolution of the data, the density at each point $\vec{r}$ in space is given by

$$\rho\left(\vec{r}|R\right) = \sum_{i}^{N} A_i \exp\left(-\frac{\left|\vec{r} - \vec{r}_i\right|^2}{2\sigma_R^2}\right) \tag{3.2}$$

Here the summation is over all N atoms indexed by i, where the amplitude of the density is given by the atom number of the element A_i, $\vec{r}$ is the position of atom i, and the spread σ_R is a function of the resolution R given by

$$\sigma\left(R\right) = \frac{1}{\sqrt{2}\pi} R \tag{3.3}$$

This definition of the resolution ensures that the amplitude of the specified resolution is at $1/e$ of the maximum in Fourier space. In order to lower the resolution of the map to a lower target resolution, the density can simply be convoluted with a Gaussian kernel as

$$\rho_{\text{target}}\left(\vec{r}\right) = G_k * \rho_{\text{init}} \tag{3.4}$$

where ρ_{init} and ρ_{target} are the initial and target density, respectively, and G_k is the Gaussian kernel with standard deviation σ_k, and $*$ is the convolution operator. The convolution of two Gaussians results in another Gaussian [98] as follows

$$G_1 * G_2 = A \exp\left[-\frac{\left|\vec{r} - \left(\vec{r}_1 + \vec{r}_2\right)\right|^2}{2\left(\sigma_1^2 + \sigma_2^2\right)}\right] \tag{3.5}$$

where G_1 and G_2 are two Gaussian functions with center $\vec{r}_1$ and $\vec{r}_2$ and width σ_1 and σ_2, and A a normalization constant of no interest here. Thus, σ_k is then simply

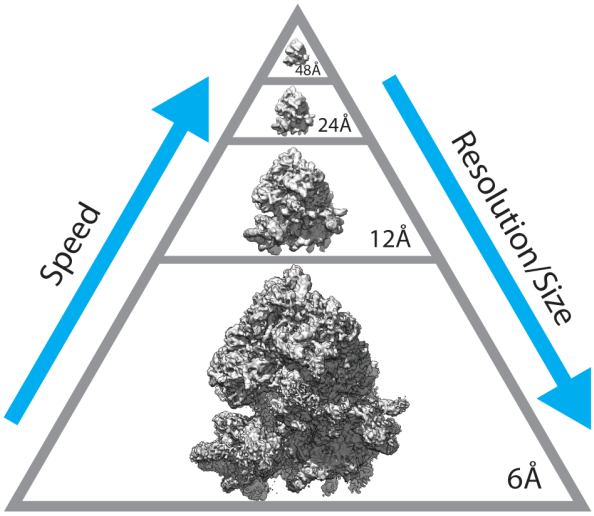

Figure 3.1 Example of a multi-scale image pyramid of the D. melanogaster ribosome (EMD-5591) . The time required for an exhaustive search increases with increasing resolution.

$$\sigma_k = \sqrt{\sigma^2_{\text{target}} - \sigma^2_{\text{init}}} \tag{3.6}$$

This procedure gives a handle and tool to lower the resolution of a map to a specified target resolution. After lowering the resolution, the data were resampled such that the voxel spacing was $1/4^{\text{th}}$ of the new resolution using simple tri-linear interpolation.

3.2.2 Leveraging the limits using multi-scale image pyramids

The rapid advancement of the cryo-EM field has resulted in an impressive increase in the number of high-resolution density maps and corresponding atomic models. The increase in the level of detail, however, also requires the number of voxels to represent the data to rise. Consequently, the time required for an exhaustive search can increase dramatically as fitting algorithms typically use the FFT for rapid translation correlation scans, which scale with $N \log N$ where N is the number of voxels.

The actual level of detail present in current high-resolution maps may, however, be far surpassing the minimal required information to unambiguously fit a subunit into the density. The superfluous amount of information can be leveraged by building a multi-scale image pyramid to

speed up the search: by progressively lowering the resolution and subsampling the data, the size of the density is reduced, which subsequently results in lower computational resources and time requirements. However, for the image-pyramid concept to work effectively, the resolution boundaries to perform a successful fitting of a particular subunit must be established. A natural parameter to investigate is the size of the subunit, expressed here simply as the number of residues, since larger chains carry more information and thus require a lower level of detail to be properly fitted in the density. Once the success rate for correct fitting of differently sized chains has been established, this information can be used to extract the required resolution for a specific chain. An image-pyramid can then be built by creating densities at different scales to fit subunits of various sizes. An example of such an image pyramid is shown in **Figure 3.1**.

3.3 Results and discussion

3.3.1 Success rate of different scoring functions

We first determined the best performing cross correlation score for fitting subunits in the experimental maps. As fine rotational searches are computationally demanding, we performed at this stage solely a translational correlation scan with the correct orientation of each chain. If no correct local cross correlation minimum can be found here, a fine search is futile. The success rate of fitting a subunit correctly is plotted against the resolution of the data in **Figure 3.2**A. A fit is considered successful if the subunit is placed within 2 voxels of the true solution, and we only considered the best-ranked fit, i.e. the fit with the highest correlation score. Congruent with an earlier analysis of noise-free simulated data, the L-CW-LCC score performs the best of the four, followed by the L-LCC, CW-LCC and LCC (see **Chapter 2**). Remarkably, the L-CW-LCC is capable of fitting about 90% of the 379 subunits unambiguously down to a resolution of 13Å. Inclusion of the Laplace pre-filter has the biggest impact and increases the applicable resolution extent by about 5Å. The core-weighted approach has a smaller impact and extends the resolutions 1 to 2Å further.

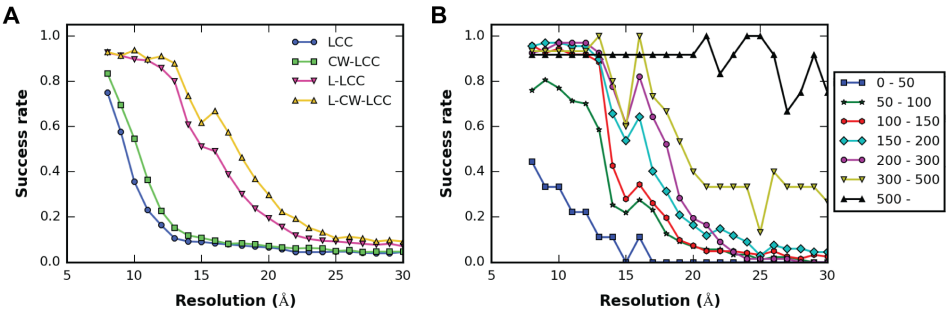

Figure 3.2 Aggregate fitting results of the 5 ribosome cases. (A) Success rate of fitting a subunit unambiguously at the correct position for four correlation scores as a function of the density map resolution. (B) Success rate of correctly fitting a subunit consisting of a given number of residues with the L-CW-LCC score as a function of the density map resolution. The subunits were divided into seven categories based on their respective number of residues.

3.3.2 Size dependence of success rate

To quantify the success rate of the L-CW-LCC further, we performed a fine rotational search (6.6° interval, 27672 orientations) for each subunit. We divided the chains in 7 categories based on their size represented by their number of residues (**Table 3.2**). The success rate of fitting each category of subunits is shown in **Figure 3.2**B. We again only considered unambiguous fits, i.e. the top ranked fit. As expected, the smallest chains have the lowest success rate. Even when fitting in 8Å resolution data the success rate is smaller than 50%. This increases to around 80% already for subunits with a residue count of 50 to 100. The success rate stabilizes to 90% for larger sized chains and is stable down to 12Å resolution data. After the 12Å point, the success rate drops rapidly, though less strongly for larger chains. For subunits larger than 500 residues, which also include the rather large rRNA chains, the success rate remains stable down to 20Å. We conclude from this analysis that the bulk of the subunits can be unambiguously fitted in the density down to 12Å resolution.

Interestingly the success rate spikes locally at 16Å resolution, and is more pronounced for larger chains. This can also be observed in **Figure 3.2**A only for the L-CW-LCC score. The reason for this is not fully apparent. It might be an artifact of the core-weighted procedure: the core-indices of subunits consisting of multiple subunits may shift suddenly and coalesce, locally increasing the sensitivity of the score. Another reason

Table 3.2 Size categories used during the analysis with the number of corresponding subunits.

Number of residues	Number of subunits
0 – 50	9
50 – 100	87
100 – 150	122
150 – 200	67
200 – 300	67
300 – 500	15
500+	12

might be that for those subunits the local resolution is significantly lower, and that fitting with a high-resolution template of the subunit results in too much noise entering the correlation score, which is remedied by filtering the template further down to lower resolutions. Although this has no impact on the main finding of the fitting analysis, the observation is intriguing.

3.3.3 Detecting over-interpreted regions of the density

The advantage of objectively fitting subunits in the density and characterizing the success rate is that it allows the identification of possibly over-interpreted regions of the density. For example, in the largest size category the elF3c chain (543 residues) of EMD-2845 was placed incorrectly in the density. When inspecting the current fit (**Figure 3.3**), we can see that the global features of the chain are present in the density, although it is not of sufficient resolution to identify the secondary structure elements, and that some parts are sticking outside the density. This was also implicated by the authors, as the local resolution of the density drops to around 10 to 15Å in that region [96]. Interestingly, the chain is unambiguously fitted when the resolution of the data is filtered to 21Å resolution, indicating that the global features of the chains are indeed consistent with the data, but the high-resolution structure might be over-interpreting the data.

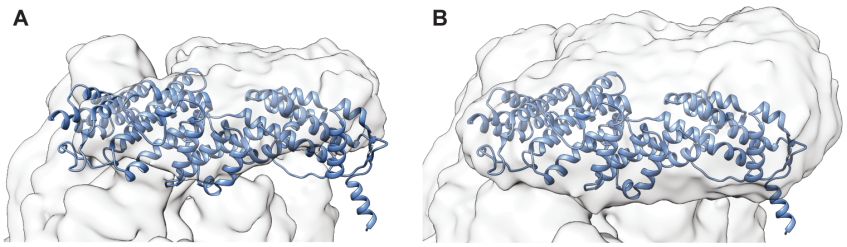

Figure 3.3 The eIF3c chains as currently placed in the density by manual rigid-body fitting (EMD-2845, 4UER) The density is shown at an iso-contour level of 0.03 (A) and 0.01 (B).

3.3.4 Fast fitting with multi-scale image pyramids

Now that resolution indicators are defined for reliably fitting a particular size chain into the density, this knowledge can be leveraged through building a multi-scale image pyramid to speed up the search. To demonstrate the speedup that can be achieved, we applied our approach to another ribosome case with a reported resolution of 5.7Å (EMD-2917, 5AKA) [99]. We constructed an image pyramid by filtering and resampling the original data down to 9, 12, 13 and 20Å resolution (**Figure 3.4**). We only fitted chains larger than 50 residues: chains consisting of 50 to 100 residues were fitted in the 9Å resolution density, chains consisting of 100 to 300 residues in the 12Å data, chains consisting of 300 to 500 residues in the 13Å map, and for subunits bigger than 500 residues we used the 20Å data.

All 31 chains could be unambiguously fitted considering only the top solution with the best cross-correlation, with the exception of the 4.5S RNA consisting of 74 residues. A local cross correlation maximum can be found at the correct location, with the successful fit placed at rank 17. This is probably due to the local resolution of the data dropping to around 10 to 12Å in that region, indicating flexibility of the chain [99]. The time required to fit one subunit into the original map ($180\times180\times180$ voxels) is approximately 10 hours using a single AMD Opteron 6320 CPU processor and 40m for an NVIDIA GTX 680 GPU. This reduces to 6h and 29m for the 9Å resolution data ($160\times160\times160$), 2h and 7m for the 12Å data ($108\times100\times120$), 1.5h and 5m for the 13Å data ($96\times96\times108$), and 20m and 1m for the 20Å data ($64\times64\times72$), respectively. Thus, the speed increase is up to 30 times for CPU and 40 times for GPU

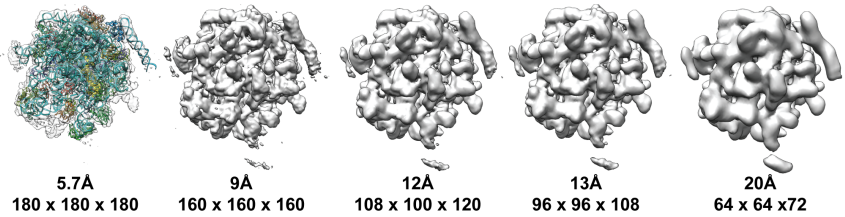

Figure 3.4 Cryo-EM data of E. coli ribosome (EMD-2917) at different resolutions with the deposited structure (5AKA) fitted into the original map (left). The resolution and the size of the data, the latter expressed in numbers of voxels, are indicated under each density.

calculations for the larger subunits, at only a small cost in the success rate of fitting.

3.4 Conclusions

Here we have explored the resolution limits of rigid body fitting in high-resolution cryo-EM densities, ranging between 5.5 and 6.9Å resolution, using 5 different ribosome cases. We have shown that also for experimental data the L-CW-LCC score is the most sensitive of the 4 correlation-based scores tested and that it can unambiguously fit most chains objectively at the correct location. In addition, we quantified the success rate of fitting subunits based on their size represented by their number of residues. As expected, larger subunits require a lower level of detail to be unambiguously fitted into the density. This phenomenon can be leveraged by building an image pyramid, i.e. representing the data at different resolutions, and subsequently fitting a subunit in the smaller, lower-resolution density dataset. The resulting speed increase can be up to 30-fold for CPUs and 40-fold for GPUs with virtually no loss in the success rate of fitting. We have implemented the use of image-pyramids in PowerFit for fast objective fitting of high-resolution structures in lower-resolution density maps.

Chapter 4

Integrative modeling of biomolecular complexes: HADDOCKing with cryo-EM data

4.1 Introduction

Protein interactions underlie most of the complexities encountered in the cell. They play a determining role in processes ranging from protein translation to muscle contraction. Numerous diseases are the result of mutations at the interface of protein complexes [100, 101]. For a thorough and fundamental understanding of these processes and rational drug design, knowledge of these interactions and interfaces at an atomic level is of paramount importance [4, 102]. Unfortunately, the number of available high-resolution structures of protein complexes determined by either X-ray crystallography or NMR spectroscopy remains rather sparse compared to the size of the interactome [103, 104].

Cryo-electron microscopy (cryo-EM) is a technique capable of imaging large biomolecular complexes in their native hydrated state [105]. The resolution is, however, usually limited to such extent that a direct atomic view of the interface is out of the question. In order to remedy this, cryo-EM data are often combined with high-resolution atomic structures [70]. The simplest and most common way of building macromolecular assemblies into cryo-EM maps is by manual fitting of atomic structures using dedicated graphics software [106, 107]. A more objective but less used method is full exhaustive search rigid body fitting, for which a plethora of software has been developed as reviewed in Esquivel-Rodríguez and Kihara [70]. Still, as the resolution decreases, placement

of subunits becomes ambiguous, and more models need to be sampled and/or additional data incorporated into the modeling to generate sensible models.

Protein-protein docking is in principle well suited for this task [9, 108], since it naturally samples a large number of conformations and can take into account additional sources of information for scoring and/or for driving the docking process [13, 14]. Several docking programs have incorporated cryo-EM data into their work flow. MultiFit automatically segments the cryo-EM density using a Gaussian mixture model to deduce anchors, subsequently docking the components of the complex onto the anchors [54]. EMLZerD uses the cryo-EM data to score the models using 3D Zernike descriptors [109]. A recent approach has been implemented in ATTRACT-EM [110], which represents the cryo-EM data by a Gaussian mixture model and fits the subunits into the map in a procedure reminiscent of Kawabata's approach [111]; the resulting models are then refined. Most of these methods, however, separate the use of the cryo-EM data from the use of other sources of information: They first fit the structures in the density and only afterwards might take into account the physico-chemical properties (energetics) of the interface. Furthermore, they usually do not actively use additional orthogonal information that may be available, such as for example mutagenesis or mass-spectrometry cross-link data.

Only few approaches have been published that can incorporate a variety of data [112], one of which is the Integrative Modeling Platform (IMP) developed in the Sali group, which has the capability of integrating among others cryo-EM data [113–115]. Another approach is our in-house data-driven docking software HADDOCK [42, 116], which is already capable of actively using information from various sources, such as mutagenesis, NMR H/D exchange and cross-links data to name only a few. In addition, it is able to deal with multiple subunits [62], can handle proteins, peptides [65], DNA [59] and RNA complexes and any combination thereof. HADDOCK leveraged its unique ability to combine multiple structural data into the modeling process to implement powerful strategies to deal with large domain conformational changes [117]. Here we describe how we have incorporated cryo-EM data into HADDOCK, such that the density is actively used as an additional energy term during docking, scoring and flexible refinement. These cryo-EM restraints can be combined with all other already available sources of

information and restraints supported in HADDOCK. We first report on the optimization and benchmarking of our method on 17 complexes from the protein-protein docking benchmark 4.0 [118] using simulated data of 10, 15 and 20Å, and a multi-component symmetrical complex. Then we demonstrate its applicability on five cases using available experimental data for two ribosome complexes, based on 9.8Å [87] and 13.5Å [88] data; two virus-antibody complexes using 8.5Å [119] and 21Å [120] resolution data; and a symmetric pentamer using 16Å negative stain data [121]. In several cases additional interface information is included based on mutagenesis data and the biology of the system. The resulting models have high quality interfaces without the clashes usually found in manually fitted models, revealing new details of the interactions.

4.2 Results and discussion

4.2.1 Implementation of cryo-EM data into HADDOCK

We first describe the implementation of cryo-EM restraints into the rigid body docking stage of HADDOCK (HADDOCK-EM, **Figure 4.1**). The approximate position of the center of mass (COM) of each chain in the density map is represented by a centroid. The positions of these centroids can be determined in multiple ways: subunits can first be placed manually in the density to the correct position after which the COM can be calculated; a full-exhaustive cross correlation search of the chains in the density using rigid body fitting software can be used (e.g. Hoang et al. [78]) to extract positions corresponding to high cross correlation values; several automatic methods have been devised for simultaneous centroid placement [54, 122–124]; a more elaborate approach combines cross-link data with the cryo-EM map to infer the positions of the subunits [125].

Once the centroids have been determined the docking can start. First the chains are separated in space at an approximate minimal distance of 25Å from each other and given a random orientation. Distance restraints are defined between the COM of each protein to either a specific centroid if one is able to distinguish the two chains in the density, or ambiguously to all centroids if the chains cannot be distinguished in the density. The former can be interpreted as unambiguous and the latter as ambiguous distance restraints. We thus transform the density data

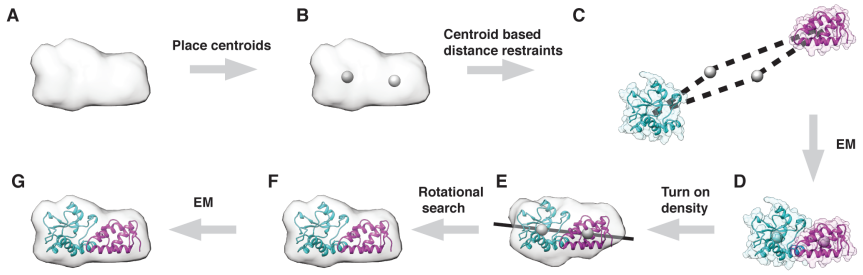

Figure 4.1 Representation of the rigid-body docking protocol in HADDOCK-EM. (A) Simulated 15Å cryo-EM data of the 7CEI complex. (B) The density with centroids (grey spheres) representing the approximate center of mass (COM) of each subunit. (C) Initial docking setup in HADDOCK. Distance restraints are defined between the COM of chain A (light-grey) and B (dark-grey) of 7CEI and their corresponding centroids. (D) An initial complex is formed after a rigid body energy minimization (EM). (E) The position of the subunits is approximately correct, but their orientation in the cryo-EM map should still be determined. (F) A fine rotational search is performed around the axis that is formed by the line joining the two centroids. The orientation with the highest cross correlation value is chosen. (G) A final rigid body EM is performed now directly against the cryo-EM data using a cross correlation-based potential without the centroid-based distance restraints.

into distance restraints for several reasons. First and foremost, this increases the radius of convergence of pulling the chains into the density towards specified positions compared to using a cross correlation potential, making the approach more robust. Indeed, when using solely the cross correlation, we found that the chains often get stuck in local minima before they can even interact with each other. Secondly, the distance restraints approach falls within the original philosophy of HADDOCK, making it easier to combine cryo-EM data with other relevant information sources. Having defined the cryo-EM derived distance restraints, we then dock the initial complex by means of rigid body energy minimization, which effectively positions it into the cryo-EM map to fit the centroids. In the case of binary complexes, the optimal orientation of the complex with respect to the density still needs to be determined since the centroid-based docking allows for rotational ambiguity. Therefore, we perform a fine rotational search of the complex around the axis formed by the line joining the centroids and score each orientation using the cross

correlation value between the model and the map. The orientation corresponding to the highest cross correlation value is further refined using a rigid body energy minimization where the energy consists of the non-bonded interaction terms of classical force fields (intermolecular van der Waals and electrostatic energies) and an added cross correlation potential. Typically 10000 solutions are generated at the rigid-body docking stage. All calculations are performed with CNS (Crystallography and NMR System) [67] (see Experimental Procedure section for details).

After the rigid body stage, the generated solutions are scored with the HADDOCK-EM-it0 score, which correspond to the original HADDOCK score (see **Eq. 4.1**) complemented with a local cross correlation-based energy (see Experimental Procedures, **Eq. 4.6 – 4.7**). The 400 best scoring models are then refined using the standard HADDOCK refinement protocol with an additional correlation-based potential to further fit the chains into the density, while reckoning with the energetics of the system.

4.2.2 Impact of cryo-EM data in the rigid body docking stage

Since the HADDOCK protocol consists of several stages (rigid body docking and scoring (it0), and flexible refinement stages in vacuum (it1) and explicit solvent (itw)), we will separately discuss the impact of incorporation of cryo-EM data on each stage. We investigated the use of 10, 15 and 20Å simulated cryo-EM data on a benchmark consisting of 17 complexes taken from the protein-protein docking benchmark 4.0 [118]. These complexes consist of 10 easy, 4 medium and 3 hard cases (based on the degree of conformational changes taking place upon complex formation) and are listed in **Table 4.1**. Even though the complexes in the benchmark are significantly smaller than what can be imaged by cryo-EM, their use is still justified to optimize our protocol and investigate the limits of using density data during the docking.

As a reference to assess the performance of using cryo-EM data in the it0-stage, we used the ab initio mode of HADDOCK (HADDOCK-CM), which uses center of mass distance restraints between molecules to drive the docking [64]. We investigated two different performance indices at this stage, namely the interfacial quality of the best-generated solution, or interface RMSD (i-RMSD) as defined by the CAPRI standards [126],

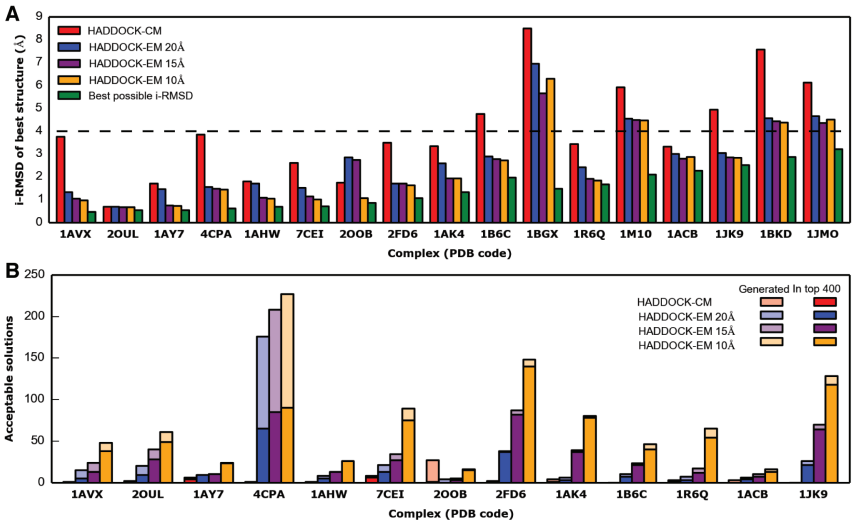

Figure 4.2 Quality and number of generated acceptable models after rigid-body docking (it0). (A) Interface RMSD (i-RMSD) of the best model generated after the it0-stage for the 17 complexes of the benchmark. White bar: HADDOCK-CM (ab initio docking mode with center-of-mass restraints); Light-grey, grey and dark-grey bar: HADDOCK-EM using 20, 15 and 10Å simulated cryo-EM data, respectively; Black bar: minimal i-RMSD of unbound compared to bound complex. The complexes are ordered according to their difficulty level. The dashed line represents the cutoff for an accetable solution (i-RMSD 4Å). (B) Number of acceptable solutions in the 10000 generated models after the it0-stage (bar height) and number of acceptable solutions in the 400 best scoring models (inner solid bar). Only complexes for which acceptable solutions were generated are displayed. Light-grey: HADDOCK-CM; Grey, dark-grey and black bar: HADDOCK-EM using 20, 15 and 10Å simulated cryo-EM data.

and, secondly, the number of acceptable solutions at the rigid-body docking stage among the 10000 models generated. We define an acceptable solution as having an i-RMSD 4.0Å from the native complex.

As can be seen in **Figure 4.2**A, HADDOCK-CM generates at the rigid body stage at least one acceptable solution out of 10000 in 11 of the 17 cases, of which 9 come from easy and 2 from medium difficulty targets. The HADDOCK-EM protocol generates at least one acceptable solution in 13 out of 17 cases, independent of the resolution of the simulated density maps used for the docking, of which all 10 easy targets, 2 medium and 1 hard target. The quality of the best-generated model improves

Table 4.1 Description of the complexes in the benchmark. The 17 protein-protein complexes used during the optimization and benchmarking of HADDOCK-EM. The complexes were taken from the protein-protein docking benchmark 4.0 [118].

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

[a] The category of the complex: E = Enzyme/Inhibitor or Enzyme/Substrate; A = Antibody/Antigen; O = Others.

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

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

for all complexes compared to HADDOCK-CM, except for the smallest 2OOB complex when using 15 and 20Å resolution data. The average i-RMSD improvement is 1.2, 1.5 and 1.6Å when using 20, 15 and 10Å data, respectively. Even for complexes for which no acceptable solutions were generated, there is a considerable increase of quality, e.g. the i-RMSD

of the hard 1JMO complex decreases from 6.13 for HADDOCK-CM to 4.66, 4.35 and 4.51Å when using 20, 15 and 10Å data, respectively.

Moreover, not only the quality of the interface of the best model benefits from the use of cryo-EM data, but the number of generated acceptable solutions also increases significantly (**Figure 4.2**B). For HADDOCK-CM the median number of generated acceptable solutions is 1, while for HADDOCK-EM it raises to 8, 17 and 46 when using 20, 15 and 10Å data, respectively. The only complex where HADDOCK-CM actually generates more acceptable solutions compared to HADDOCK-EM is again the small globular 2OOB complex.

As our protocol is dependent on the input of centroid coordinates, we also investigated its sensitivity to incorrect centroid placement. To this end, we repeated the docking for 5 cases where both centroids were separately displaced by 3, 5 and 7Å in a random direction. The total error in placement was thus 14Å total in the latter case. The difference in the number of acceptable solutions generated in the top 400 differed per case (see **Table S4.2**). Only at 7Å displacement of both centroids does the number of acceptable solutions in the top 400 decrease consistently, but is still significantly larger compared to HADDOCK-CM. Thus, our approach is robust against centroid placement errors up to at least 7Å.

4.2.3 Impact of cryo-EM data on the scoring of rigid body docking solutions

To incorporate the cryo-EM data into the scoring function, we supplemented the original HADDOCK score with a local cross correlation (LCC) energy term (HADDOCK-EM score). The efficiency of this combined score is shown in **Figure 4.2**B. The HADDOCK-CM models were scored with the original HADDOCK score, which resulted in at least one acceptable solution in the top 400 models for 7 out of the 11 successful cases, where at least one acceptable solution was generated. The HADDOCK-EM models were scored with the HADDOCK-EM score, which resulted in at least one acceptable for all 13 successful cases, irrespective of resolution, with the exception of the 2OOB complex using 20Å resolution data.

To investigate the effect of the LCC term in the HADDOCK score, the total number of acceptable solutions in the top 400 was calculated for

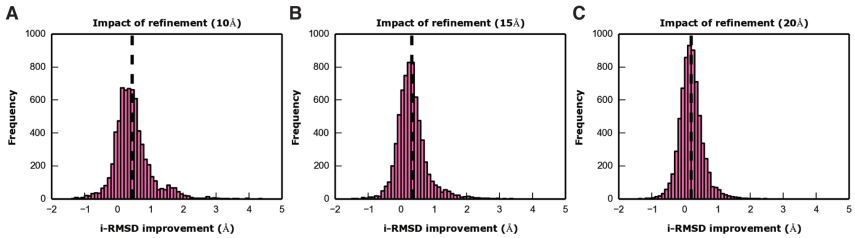

Figure 4.3 Effect of the flexible refinement stage with cryo-EM restraints on i-RMSD. The i-RMSD improvement (i-RMSD it0 – i-RMSD itw) for all refined complexes after itw when using 10 (A), 15 (B), and 20Å (C) data plotted as a histogram. Positive values indicate a decrease in i-RMSD toward the native structure. The dashed vertical line in the figures represents the average i-RMSD improvement.

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

The discriminative ability of the LCC-term increases with the resolution, as expected. When plotting the LCC versus the i-RMSD (see **Figure S4.1**), we observe a funnel shape for most complexes, with high LCC values found for complexes with low i-RMSD values. This becomes even more pronounced as the resolution of the data increases. For higher i-RMSD the correlation is lost and the LCC is no longer indicative of the quality of the solutions as was observed before [127]. It should further be noted that the absolute value of the LCC-term is not indicative of the quality of the model. For example, when using 20Å data correlation values of > 0.9 are routinely found for non-native models. As such, the correlation value only has meaning in a comparative setting, urging the need to sample and score multiple conformations.

4.2.4 Effect of cryo-EM data on the flexible refinement stage

Next we investigated the impact of incorporating cryo-EM restraints on the flexible refinement stage of HADDOCK. We calculated the i-

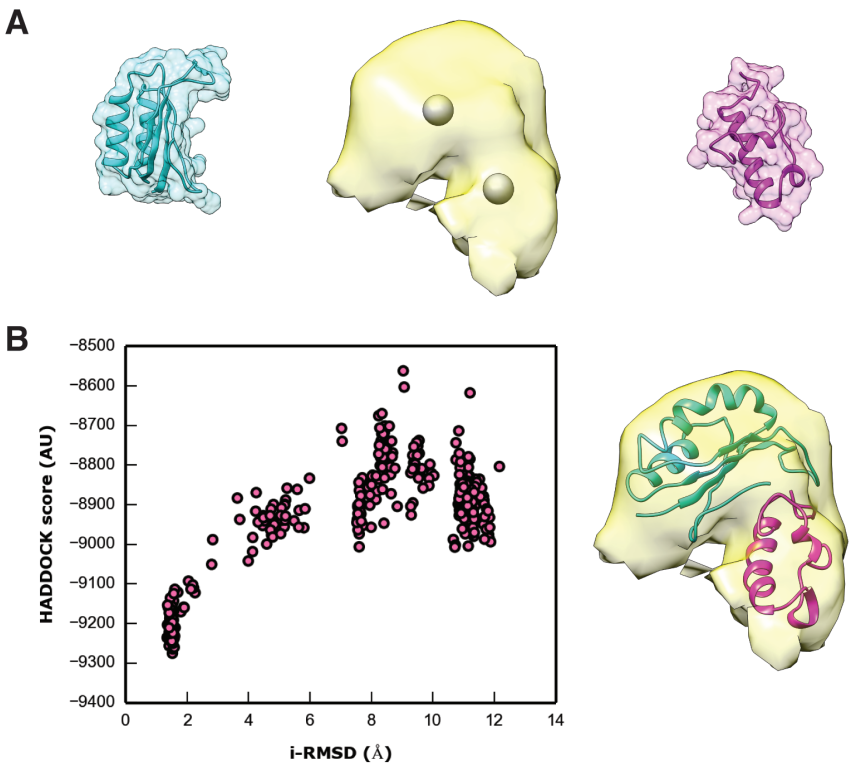

Figure 4.4 Cryo-EM driven HADDOCKing docking of the ribosomal proteins S7 and S19 onto the 30S E. coli ribosome. (A) Docking setup used in HADDOCK showing the S7 (cyan) and S19 (magenta) protein, the centroids and the density (yellow). (B) The HADDOCK-EM score of the 400 refined models plotted versus their i-RMSD from the 2YKR-structure. Next to it the solution with the best HADDOCK score and an i-RMSD of 1.56Å is shown in the cryo-EM density.

RMSD improvement of the 400 best scoring it0-models after each refinement stage for all complexes. A histogram of i-RMSD improvements after the it0 and itw stage is shown in **Figure 4.3**. The average i-RMSD improvement after refinement when using 20Å data is 0.20Å with a maximum of 2.49Å. This increases to an average of 0.33 and 0.45Å and a maximum of 3.34 and 4.37Å when using 15 and 10Å data, respectively. The average i-RMSD improvements between it1 and itw are modest: 0.04, 0.05 and 0.10Å when using 20, 15 and 10Å data with maximums of 0.25, 0.34 and 0.43Å, see **Figure S4.2**. So the bulk of the improvement is gained during the it1-stage, which was also previously noted (see Figure 2

in [43]). The maximal improvement observed with cryo-EM restraints is about two times larger than what was previously observed in an analysis of our CAPRI predictions. This substantial improvement is also reflected in the increased number of acceptable solutions after the refinement for each complex (**Table S4.1**). The number of cases with at least one acceptable increases from 13 for 20Å resolution data to 15 for the 15 and 10Å resolution data (**Table 4.2**). The resulting models are ultimately re-scored using the itw-HADDOCK-EM score (**Eq. 4.7**). The enrichment of models in the top 400, 10 and 1 compared to HADDOCK-CM are given in **Table S4.3**.

4.2.5 Docking two ribosomal proteins using experimental 9.8Å cryo-EM data

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

We docked the two proteins using only the fraction of the cryo-EM density that can be attributed to the two proteins (**Figure 4.4**A). The centroids were determined by calculating the position of the COM of each protein as they were currently placed in the density. Applying HADDOCK-EM resulted in 15 clusters, with the best scoring cluster containing 105 of the 400 generated solutions of which the best scoring complex has an i-RMSD of 1.56Å compared to the crystal structure (**Figure 4.4**B).

4.2.6 Integrative modeling of KsgA with rRNA using 13.5Å cryo-EM data

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

Table 4.2 i-RMSD values of the best generated complex after the rigid-body docking (it0) and final water refinement (itw) stages. The quality in terms of i-RMSD values of the best solution generated after it0 and itw stages is given for each complex at the three cryo-EM density resolutions. The i-RMSD is calculated by fitting the solution on the backbone atoms of the residues involved in intermolecular contacts in the native complex within a cutoff of 10Å.

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

rRNA are involved in the interaction (**Figure 4.5**A), which has been corroborated by hydroxyl radical foot-printing data [128]. Mutagenesis data show that the positively charged residues R221, R222 and K223 of KsgA are important in the interaction (**Figure 4.5**B) [88].
The 13.5Å cryo-EM map has a corresponding current PDB model (4ADV). This model, however, contains a large number of clashes at the interface (>100). Furthermore, it reveals no favorable interactions and fails to

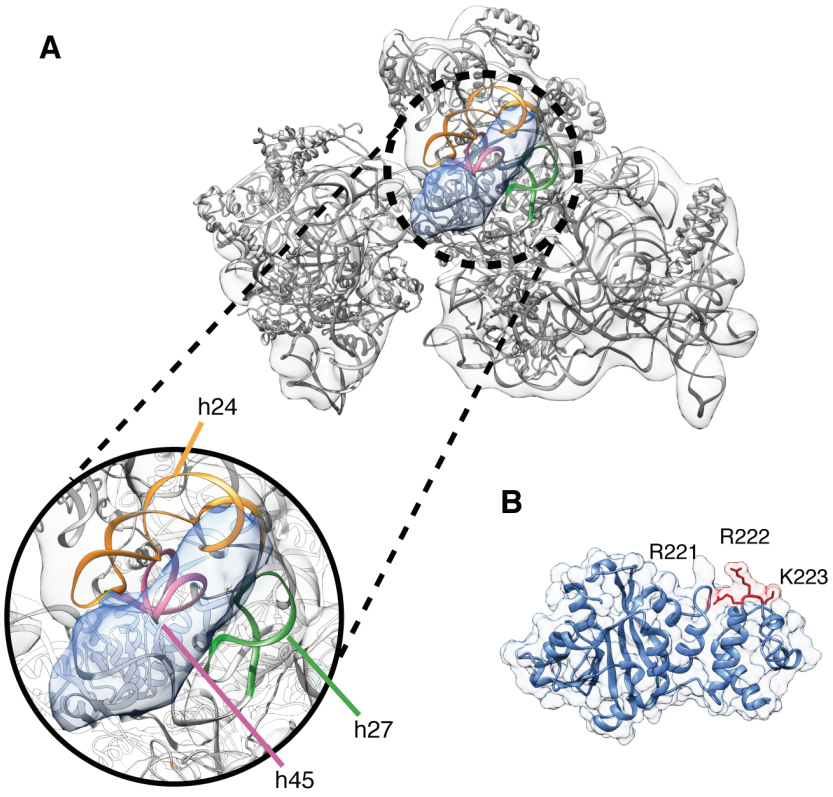

Figure 4.5 Cryo-EM and mutagenesis data of the 30S maturing E. coli ribosome and its current model. (A) The 13.5Å cryo-EM map of the maturing 30S E. coli ribosome with its current PDB model fitted inside. The density of KsgA is shown in blue, and the helices of the rRNA are shown in orange (h24), green (h27) and pink (h45). The binding site of KsgA is shown below enlarged. (B) Crystal structure of KsgA of E. coli with the three key residues shown in red.

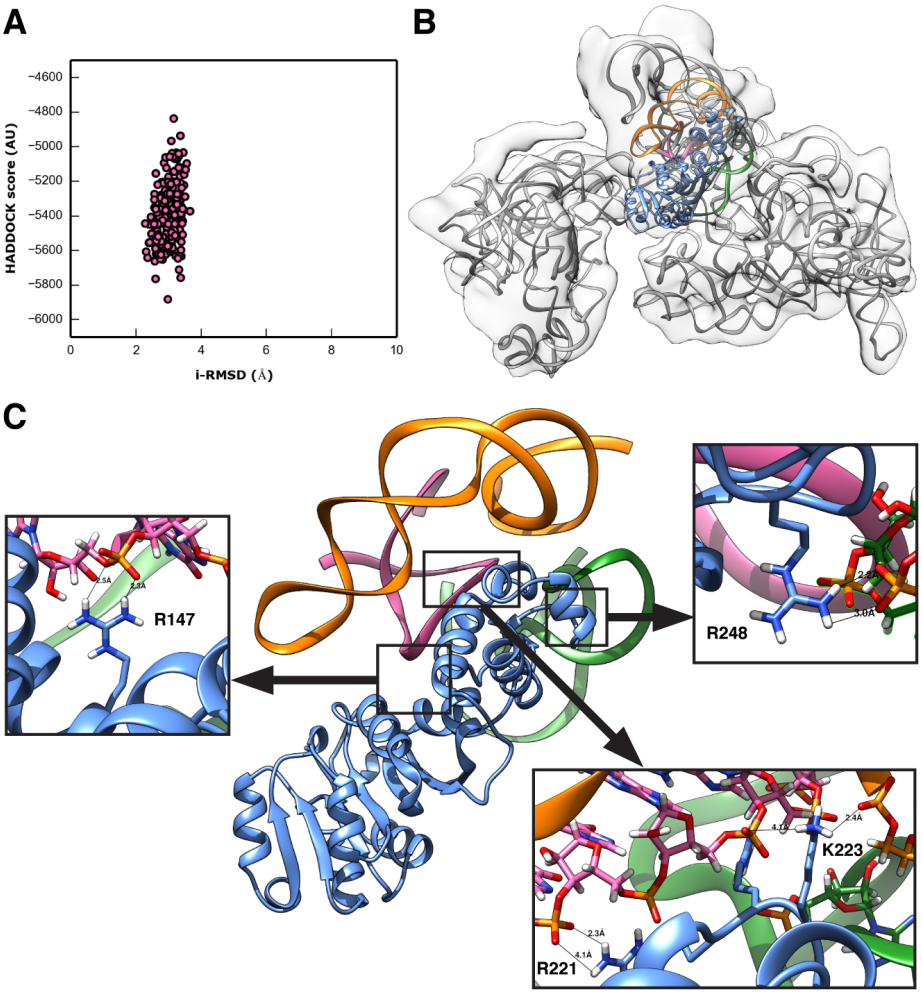

Figure 4.6 Cryo-EM driven HADDOCKing of KsgA ontop the 16S rRNA of E. coli. (A) The HADDOCK-EM score of the 400 refined models plotted versus the i-RMSD compared to the 4adv-model. (B) Binding mode of the best scoring HADDOCK-EM model, together with the 13.5Å cryo-EM map. (C) Close up of the binding of KsgA with the rRNA. The right bottom figure shows the favorable hydrogen bonds formed by the three key residues R221, R222 and K223. At the left and upper right side the additional evolutionary conserved residues R147 and R248 are shown forming favorable hydrogen bonds with the backbone of the rRNA.

give a clear explanation for the importance of the arginine residues identified by mutagenesis (**Figure S4.3**). This is a typical side effect from manual rigid body fitting. Running HADDOCK-EM using the radical foot-printing, mutagenesis and cryo-EM data results in a single cluster (**Figure 4.6**A) of which the best solution has an i-RMSD of 2.8Å compared to the 4ADV-model. The placement and orientations of the chains in the density are similar to the rigid body fitted model as defined by the cryo-EM data. The HADDOCK-EM model is, however, of much better quality: it contains no clashes and reveals favorable hydrogen bonds made by R221, R222 and K223 with the backbone of the rRNA. Moreover, new potentially key residues can be identified, such as R147 and R248 (**Figure 4.6**B). Coincidentally, these newly identified residues are also highly conserved, corroborating our docking results (see **Figure S4.4**).

4.2.7 Modeling virus-antibody complexes using 8.5Å and 21Å cryo-EM data

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

For both cases we performed a HADDOCK run, combining the cryo-EM data with interface information. Since the binding regions on the antibody are known as well as the virus capsid proteins, residues that were within 5Å of the other chain in the deposited atomic models were used as active residues. The solutions of the adeno-associated virus 2 converge into 1 cluster with an i-RMSD less than 1.5Å from the deposited model (**Figure 4.7**A). However, when zooming in on the interface of the best scoring HADDOCK model, the interactions show an extensive hydrogen bond network between the envelope protein and the antibody in contrast to the deposited model (**Figure 4.7**B).
The HADDOCK solutions of the Dengue virus cluster into two groups, with an approximate i-RMSD of 2.0Å and 4.5Å with respect to the deposited model (**Figure 4.7**C). Inspecting the interface of the best scoring HADDOCK model again shows favorable interactions between the prM

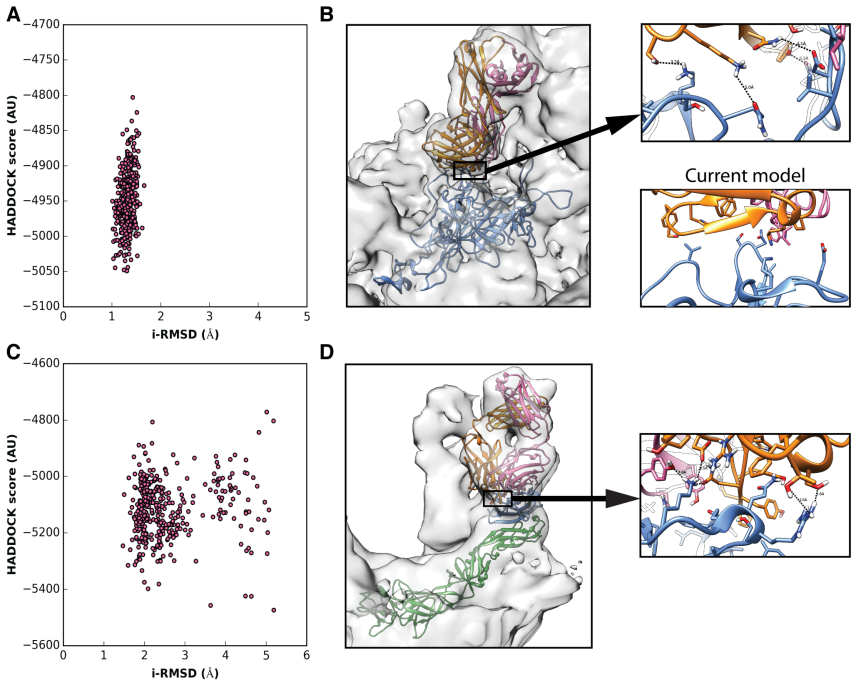

Figure 4.7 Virus-antibody HADDOCKing using 8.5Å and 21Å cryo-EM data. (A) The HADDOCK-EM score of the 400 generated models of the adeno-associated virus 2-antibody complex versus their i-RMSD using 3J1S as a reference. (B) Best scoring HADDOCK model shown in the cryo-EM density. The envelope protein (blue) forms favorable interactions with the antibody A20 chains (orange and pink). The 3J1S interface is shown under the interface close up. (C) The HADDOCK-EM score of the 400 generated models of the Dengue virus-antibody complex versus their i-RMSD using 3J42 as a reference results in two clusters. (D) Best scoring HADDOCK model shown in the density. The Dengue envelope protein (green) with the prM protein (blue) forms favorable interactions with the 2H2 Fab-fragment (orange and pink).

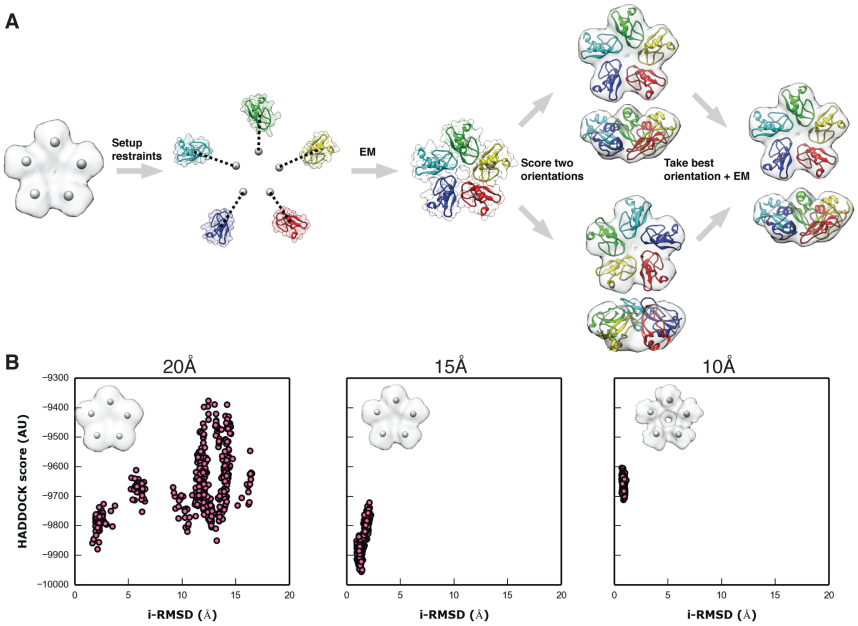

Figure 4.8 HADDOCK-EM with symmetry protocol applied on the trypsin inhibitor and large terminase pentamer. (A) Protocol of HADDOCK-EM with symmetry during the rigid body stage. After determining the centroids in the density, each subunit is placed on a circle concentric with the centroids' midpoint. C5-symmetry is imposed on the system from the beginning and ambiguous distance restraints are generated between the center of mass of each subunit and each centroid. An initial complex is formed by rigid body energy minimization (EM). To orient the complex properly in the density, we calculate the cross correlation of two orientations of the complex with the cryo-EM data. A second round of rigid body EM is performed on the orientation with the highest cross correlation directly against the cryo-EM data. (B) The HADDOCK-EM score versus the interface-RMSD compared to the native complex (1B0C) are plotted for the 400 refinement complexes using 20, 15 and 10Å simulated data.

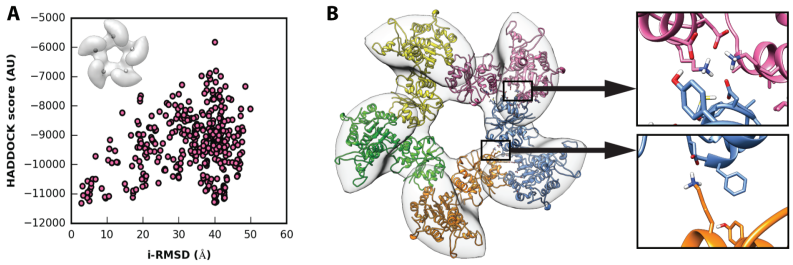

Figure 4.9 HADDOCK-EM results on the large terminase complex. (A) The HADDOCK-EM score of the 400 generated large terminase models versus the interface-RMSD. The 16Å negative stain data with centroids are shown in the left corner. (B) Best scoring HADDOCK model shown in the density. Close-ups of the interface are displayed to the right.

protein of the Dengue virus with the antibody, while the 3J42-model lacks side-chains and shows a backbone clash (**Figure 4.7**D).

4.2.8 Symmetrical multibody docking with cryo-EM data

HADDOCK is capable of using symmetry restraints to drive the docking of symmetrical assemblies. In order to combine symmetry and cryo-EM restraints the rigid body docking protocol was slightly modified compared to non-symmetric complexes, with as main difference the initial placement of the subunits (**Figure 4.8**A, Experimental Procedures). We tested HADDOCK-EM with symmetry on the cyclic pentamer of the trypsin inhibitor (1B0C). The ab initio mode of HADDOCK with C5 symmetry restraints results in two acceptable solutions after the refinement stage, with the best solution having an i-RMSD of 3.2Å compared to the 1B0C-structure. Adding cryo-EM data results in an increased number of acceptable solutions of 54, 400 and 400 when using 20, 15 and 10Å resolution data, respectively, with the best models having i-RMSDs of 1.6, 1.0 and 0.7Å (**Figure 4.8**B). Using higher resolution data also results in more compact clusters, i.e. the distribution of i-RMSD values is reduced. When using 10Å data only a single near-native cluster is observed. At 20Å resolution, multiple clusters appear and require the HADDOCK-EM score to discriminate the near-native cluster, which indeed has the best (lowest) HADDOCK-EM score.

We applied the symmetrical HADDOCK-EM protocol to model the pentameric large terminase complex of bacteriophage T7 using 16Å negative stain EM data (EMD-2355, [121]). As in the previous cases, the corresponding deposited model (4BIJ) shows clashes at the interfaces (**Figure S4.9**). The 400 generated HADDOCK models resulted in 33 clusters, with the best scoring cluster having an i-RMSD of 2.9Å compared to the 4BIJ-model (**Figure 4.9**). Again, the interface of the best scoring HADDOCK model alleviates the clashes and shows favorable interactions, while agreeing with the general binding mode of the 4BIJ-model.

4.3 Conclusion

We have fully integrated cryo-EM data into HADDOCK, allowing the direct combination of cryo-EM data with all other available sources of information that HADDOCK supports, including symmetry and ambiguous interaction restraints. The performance of this integrative docking protocol was demonstrated using simulated cryo-EM data for a benchmark of 17 non-redundant protein-protein complexes: Including the cryo-EM data into the docking significantly increases both the quality and quantity of acceptable solutions, with higher resolution data having a larger impact. Its applicability was demonstrated on two ribosome, two virus-antibody and a symmetrical case using experimental data ranging from 8.5 to 21Å. The integration of cryo-EM data with a proper physics-based force field and all other available information sources provides a powerful and user-friendly tool to generate high quality, high resolution models of macromolecular assemblies.

4.4 Experimental procedures

4.4.1 HADDOCK protocol

HADDOCK has been described in details in previous work [42, 116]. Its docking protocol consists of three stages: an initial rigid body docking stage (it0), a semi-flexible refinement stage using simulated annealing in torsion-angle space (it1) and a final flexible refinement stage in explicit water (itw). In it0, the subunits are treated as rigid entities. They are

separated in space by an approximate minimum distance of 25Å, each subunit being given a random orientation and random translation within a 10Å box. In the ab initio mode of HADDOCK (HADDOCK-CM), center of mass restraints are defined between the subunits to drive the docking, as described in Karaca and Bonvin [64]. The initial complexes are generated by rigid body energy minimization where the energy is a linear combination of the intermolecular van der Waals and electrostatic energies and the empirical distance (and other, e.g. symmetry) restraint term. Typically 10000 models are written to disk at this stage. The top 400 best scoring models are refined in it1, using multiple cycles of simulated annealing in torsion angle space. In the final itw stage, the 400 structures are refined further using molecular dynamics in Cartesian space with the complex solvated in an 8Å shell of explicit TIP3P water. Finally, the 400 structures are scored with the itw-HADDOCK score.

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

$$E_{\text{it0}} = 0.1 \cdot E_{\text{vdW}} + 1.0 \cdot E_{\text{elec}} + 0.01 \cdot E_{\text{AIR}} + 1.0 \cdot E_{\text{desolv}} - 0.01 \cdot \text{BSA} \quad (4.1)$$

$$E_{\text{it1}} = 1.0 \cdot E_{\text{vdW}} + 0.2 \cdot E_{\text{elec}} + 0.1 \cdot E_{\text{AIR}} + 1.0 \cdot E_{\text{desolv}} - 0.01 \cdot \text{BSA} \quad (4.2)$$

$$E_{\text{itw}} = 1.0 \cdot E_{\text{vdW}} + 0.2 \cdot E_{\text{elec}} + 0.1 \cdot E_{\text{AIR}} + 1.0 \cdot E_{\text{desolv}} \quad (4.3)$$

where E_{it0}, E_{it1} and E_{itw} are the scoring functions after the it0, it1 and itw stage, respectively, E_{vdW} the intermolecular van der Waals energy, E_{elec} the intermolecular electrostatic energy, E_{AIR} the ambiguous interaction restraints energy, E_{desolv} an empirical desolvation energy [129] and BSA is the buried surface area in Å^2. The energies are calculated with an 8.5Å cutoff based on OPLS parameters [130].

4.4.2 HADDOCK-EM protocol

As HADDOCK uses CNS (Crystallography and NMR System) [67] as computational engine, all crystallographic tools and energy function available in CNS are available to HADDOCK. So the cryo-EM data, represented by a 3D real scalar field, can be directly read into the CNS framework and specific energy functions, typically in reciprocal space, be used and applied. The HADDOCK-EM protocol uses in particular the

xref energy term in CNS. It is very similar to the original HADDOCK method with some adjustments mainly in it0. A graphical representation of the adjusted it0-protocol is given in Figure 1. An integral part of our protocol is the use of centroids, where each centroid represents the approximate position of the COM of a subunit in the density map. When the resolution of the cryo-EM data decreases, the orientation of the subunits can be ambiguous but the approximate placement can still be determined. This is obvious in cases where several density maps are obtained with some subunits being alternately present and absent in the set, such as in the case of the ribosome [128]. The position of the centroids can be determined in multiple ways. An objective way is to perform a full-exhaustive cross correlation search to deduce regions of high cross correlation values; the centroid can then be placed on the position with the highest value. They can be placed manually using graphics software: for example UCSF Chimera has an option to place centroids in high-density regions in the map. Another option is to place an atomic structure in the density at an approximately correct position, calculate its COM and use this as the position of the centroid. Methods for automatic simultaneous detection of centroids have also been reported [54, 122–124]. A more elaborate approach uses experimental data in conjunction with the cryo-EM map to infer the positions of the subunits, as was shown on the RNA polymerase II [125]. The centroids are entered into HADDOCK-EM as Cartesian coordinates in the start parameters. Together with the cryo-EM map and its resolution, they represent all the input required to run HADDOCK-EM.

During the docking, each subunit is given a random orientation and initially placed on a sphere centered on the midpoint of the centroids. In the case of two chains, the subunits are placed opposite each other on the sphere with a minimal distance of 25Å between them. Afterwards, for each docking trial, they are given a random rotation and translation within a 10Å box to enhance the sampling. Distance restraints are defined between the COM of each subunit and either all determined centroids as ambiguous restraints in cases where the placement of the subunit in the density is ambiguous, or a specific centroid if the placement is unambiguous. The distance restraint is described by a soft square potential between two pseudo atoms, one of which corresponding to the centroid and the other to the COM of the subunit. An initial complex is formed by rigid body energy minimization, where the energy is a combination of

the force field, the centroid-based distance restraints and other possible experimentally based distance and orientation restraints.

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

$$E_{\text{vector}} = \frac{\sum_H (F_{em} - F_c)^2}{\sum_H F_{em}^2} \tag{4.4}$$

where the summation is over all the Miller indices H up to the specified resolution of the cryo-EM map, and F_{em} and F_c are the complex-valued Fourier coefficients of the cryo-EM map and the calculated density, respectively. It should be noted that minimizing the vector residual in reciprocal space is mathematically the same as maximizing the cross correlation in real space [131] and thus we refer to this potential simply as the cross correlation. The complex is reoriented in the density conforming to the optimal cross correlation value found during the search. A final rigid body energy minimization is performed directly against the map using the cross correlation (vector potential energy term in CNS), van der Waals and electrostatic energy terms. For each complex typically 10000 models are generated this way.

The models are then scored by adding a local cross correlation (LCC) term to the regular HADDOCK score (**Eq. 4.1 – 4.3**), where the LCC is given by

$$\text{LCC} = \frac{\sum_{\text{i}} (\rho_{\text{em}} - \bar{\rho}_{\text{em}}) \cdot (\rho_{\text{c}} - \bar{\rho}_{\text{c}})}{\sigma_{\text{em}} \sigma_{\text{c}}} \tag{4.5}$$

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

$$E_{\text{it0,EM}} = E_{\text{it0}} - w_{\text{it0}} \cdot \text{LCC} \tag{4.6}$$

$$E_{\text{itw,EM}} = E_{\text{it1}} - w_{\text{itw}} \cdot \text{LCC} \tag{4.7}$$

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

It should be noted that the maximum number of subunits that can be docked simultaneously currently is restricted to 6 (this limitation will be lifted in a future version). Furthermore, in order to use the HADDOCK-EM protocol, approximate knowledge of the position of each subunit in the form of centroids is a requisite for a successful docking run. Other minor requirements are that the number of voxels in each dimension of the cryo-EM data is a multiple of 2, 3 and 5 to calculate FFTs used in the cross-correlation potential, and that the density should be converted to CNS/XPLOR format. For the latter two tasks, Python scripts are included in the HADDOCK distribution. Finally, the time required for a HADDOCK-EM run decreases with decreasing map size, since this speeds up the calculations of the FFTs.

4.4.3 Generation of simulated cryo-EM maps

For the generation of the simulated cryo-EM maps we wrote a Python script based on the molmap function in UCSF Chimera. The resulting density is described by:

$$\rho\left(\vec{r}\right) = \sum_{i}^{N} A_i \cdot \exp\left(-\frac{\left|\vec{r} - \vec{r}_i\right|^2}{2\sigma^2}\right) \tag{4.8}$$

where ρ is the density value at position $\vec{r}$; the summation is over all the atoms N; A_i is the atom number of atom i; $\vec{r}_i$ is the position of atom i; and σ is the standard deviation given by $\frac{1}{\pi\sqrt{2}} \cdot R$, where R is the resolution. This definition of the standard deviation ensures that the magnitudes of the Fourier coefficients are at $1/e$ of its maximum value at the specified resolution. The extent of the Gaussian kernel was four times the standard deviation and the voxel spacing one-fourth of the

resolution. An example of a resulting 10, 15 and 20Å map is given in **Figure S4.5**.

4.4.4 Optimizing and benchmarking HADDOCK-EM

The HADDOCK-EM protocol relies on the optimization of two parameters for the docking, namely the force constant for the centroid based distance restraints and the weight for the cross correlation energy term. In addition, the weight factors of the LCC term in the it0 and itw-HADDOCK score need to be determined. For this, we used a benchmark consisting of 17 complexes taken from the protein-protein docking benchmark 4.0 [118] (see **Table 4.1**). Centroids were determined by calculating the COM of each unbound chain that is optimally superimposed onto the native complex.

We first determined the centroid based force constant by running the benchmark at different values for the force constant, creating 10000 models for each complex in it0. Since the force constant is only used in it0, the structures were not scored nor refined. The value for the force constant that gave the most acceptable solutions, where an acceptable solution is defined as having an i-RMSD <= 4.0Å compared to the native complex, was chosen (results not shown). The i-RMSDs were calculated using ProFitV3.1 [132]. For the determination of the weight factor for the cross correlation term we followed the same protocol but with the optimized force constant for the centroid based distance restraints using simulated data at 10, 15 and 20Å which were generated as described above. This gave a value of 50 for the force constant and a weight factor of 15000 for the cross correlation based energy term, independent of the resolution (results not shown).

The weight factor for the LCC in the it0-HADDOCK-EM score was determined by running the benchmark using the optimized parameters and varying the LCC weight in order to maximize the number of acceptable solutions in the top 400 at the three resolutions of 10, 15 and 20Å. This gave a value of -400. The LCC weight factor in the itw score was determined by maximizing the number of acceptable solutions in the top 20, which gave a weight factor of -10000.

To investigate the sensitivity of the protocol to incorrectly placed centroids, we ran 5 cases of the benchmark with displaced centroids. Each centroid was moved in a random direction by taking a random point on

the unit sphere with a displacement of 3, 5 and 7Å. The solutions were analyzed as explained above.

4.4.5 HADDOCK-EM with symmetry

To leverage Cn-symmetry in cyclical symmetric complexes a few adjustments were made to the non-symmetric HADDOCK-EM protocol (**Figure 4.8**A). The main difference is in the initial placement of the subunits in the it0-stage. Instead of placing the subunits on a sphere, we place them on a circle with its center placed on the middle-point of the centroids and parallel to the plane of the centroids. The radius of the circle is chosen such that the minimal distance between two subunits is at least 25Å. The requested Cn- symmetry is imposed on the system from the start.

After the initial placement, ambiguous centroid based distance restraints are generated, i.e. we create a distance restraint between the COM of each subunit to each centroid. We form an initial complex again by performing a rigid body energy minimization, where the energy includes the force field, the centroid-based distance restraints and the already in HADDOCK available symmetry restraints. Once the initial complex is formed, it needs to be properly oriented in the density. Only two orientations need to be sampled for this, namely the current orientation and the upside-down complex. The orientation corresponding to the highest cross correlation with the cryo-EM data is chosen. A final rigid body energy minimization is performed against the map, using the cross correlation potential in combination with the force field and symmetry restraints. Typically 10000 models are generated. They are scored with the it0-HADDOCK-EM and 400 models are refined in the it1 and itw stage. The refinement protocol is similar to the non-symmetric HADDOCK-EM protocol, but with added symmetry restraints.

4.4.6 Modeling protein S7 and S19 of the 30S E. coli ribosome

For the 30S E. coli ribosome the cryo-EM data were downloaded from the EMDB (EMD-1884) with its corresponding fitted PDB file (2YKR). The density belonging to the interacting ribosomal proteins S7 and S19, was masked out as follows. First we subtracted the density of the 16S rRNA modeled at the specified resolution of 9.8Å from the

map. The density belonging to the S7 and S19 proteins was masked out by creating a binary mask around the two proteins with a shell of 7Å. The centroids were determined by calculating the COM of each chain as they were fitted in the density in the 2YKR-model. We followed the standard HADDOCK-EM protocol as described above using ambiguous centroid distance restraints. During the docking 10000 models were generated in it0 and the top 400 best scoring solutions were refined in it1 and itw. The solutions were clustered using a cutoff of 7.5Å and the i-RMSDs were calculated against the 2YKR-model, using ProFit.

4.4.7 Modeling the interaction of KsgA with the 16S rRNA of the 30S E. coli ribosome

In order to model the interaction of KsgA with the 16S rRNA of the 30S E. coli ribosome, we downloaded the 13.5Å cryo-EM map from the EMDB (EMD-2017) with its corresponding fitted PDB file (4ADV). Since the 16S rRNA chain can be unambiguously placed in the cryo-EM map, we used the 16S rRNA as it was fitted and kept the chain fixed during the it0 stage. The centroid for KsgA was determined by performing a full-exhaustive local cross correlation search using a local version of software similar as described by Hoang et al. [78], with an angular sampling interval of 5°. The resulting local cross correlation map is shown in **Figure S4.6**. The centroid was placed at the position in the cryo-EM map with the highest local cross correlation. We created an initial setup for the rigid body docking by manually placing KsgA at an approximate distance of 25Å away from the interaction surface of the 16S rRNA. During the generation of each solution in it0, KsgA was given a random orientation and a random translation in a 10Å box. The initial docking setup with the determined centroid is shown in **Figure S4.7**.

Mutagenesis data shows that the residues R221, R222 and K223 of KsgA are vital in the binding of KsgA to the 16S rRNA, and hydroxyl radical footprinting and the cryo-EM map show that the helices 24, 27 and 45 of the 16S rRNA are involved in the binding. As such, the residues 221 to 223 of KsgA and residues 768 – 773, 781, 782, 801 – 803, 899 – 902, 1512 – 1516 and 1523 of the 16S rRNA were considered active residues in HADDOCK (**Figure S4.7**C).

We generated 10000 models in the rigid body docking stage. The top 400 scoring models were only refined in it1. The water refinement stage

was skipped since full molecular dynamics simulations for 400 models with the ribosome are computationally too expensive, and the impact of the itw stage is only marginal. During the refinement, additional unambiguous distance restraints were used to keep helix-45 in its place, since it was disconnected from the main chain but does take part in the interaction. The i-RMSDs of the refined models were calculated against the current 4ADV-model.

4.4.8 Modeling the adeno-associated virus-2 complexed with antibody

For the adeno-associated virus-2 in complex with antibody A20 the 8.5Å resolution cryo-EM data were obtained from the EMDB (EMD-5424) together with the fitted PDB (3J1S). As the resolution allows an unambiguous rigid body fit of both the capsid protein and the anti-body, the centroids were determined by calculating the COM of each chain in the 3J1S-model. Residues within 5Å distance of the other interacting chain were used as active residues in HADDOCK. To speed-up the calculation, the density within a 50Å shell was masked out. I-RMSDs were calculated using 3J1S as a reference.

4.4.9 Modeling the interaction between Dengue virus and 2H2 antibody

Cryo-EM data of 21Å resolution were downloaded from the EMDB (EMD-5674) and its current model (3J42). For docking, the antibody was taken from the current model, while for the envelope-prM complex the original model was used (3C5X), since the envelope protein-prM complex in 3J42 did not have any side-chains. Centroids were determined using a full-cross correlation search with Laplace pre-filter for both subunits (see **Figure S4.8**), similar to the KsgA-ribosome case. Residues within 5Å of the other chain in the 3J42 model were used as active residues during the docking. The i-RMSDs were calculated using the 3J42 model as a reference.

4.4.10 Modeling the large terminase complex

The 16Å resolution negative stain data were obtained from the EMDB (EMD-2355) together with its deposited model (4BIJ). Centroids were determined by calculating the COM of each subunit in the 4BIJ-model. The HADDOCK-EM with symmetry protocol was used, specifically using C5-symmetry restraints. 10000/400/400 models were generated in the it0/it1/itw stage. The 4BIJ-model was used as a reference for calculating the i-RMSDs.

Notes

This Chapter is based on: G.C.P. van Zundert, A.S.J. Melquiond and A.M.J.J. Bonvin. Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. *Structure* **23**, 949–960 (2015).

Supplementary information

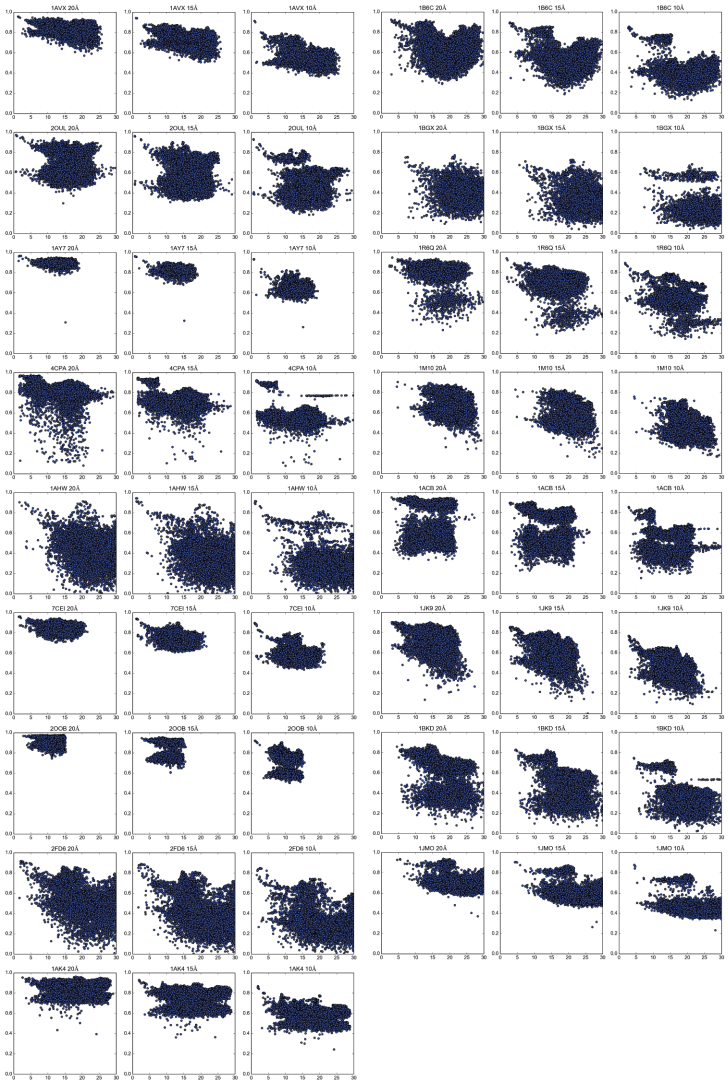

Figure S4.1 Local cross correlation scores for all 17 complexes using simulated cryo-EM data. The local cross correlation score is plotted against the i-RMSD compared to the native complex for all 17 complexes using simulated 20 (left), 15 (middle) and 10Å (right) resolution cryo-EM data.

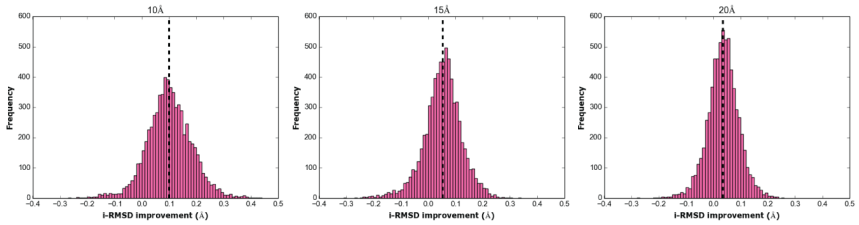

Figure S4.2 Effect of the itw-flexible refinement stage with cryo-EM restraints on i-RMSD. The i-RMSD improvement (i-RMSD it1 – i-RMSD itw) for all refined complexes after itw when using 10 (A), 15 (B) and 20Å (C) resolution data. The dashed vertical line in each figure represents the average i-RMSD improvement.

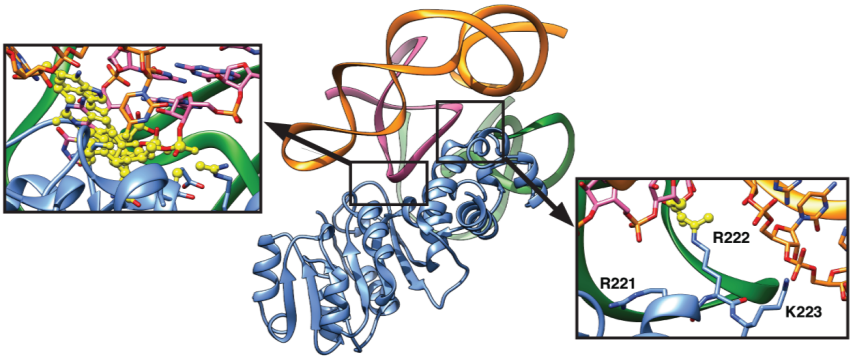

Figure S4.3 The deposited model of EMD-2017. A ribbon representation of the 4ADV-model is shown in the middle. The left and right figures are close ups of the interface. Atoms displayed as yellow balls are clashes.

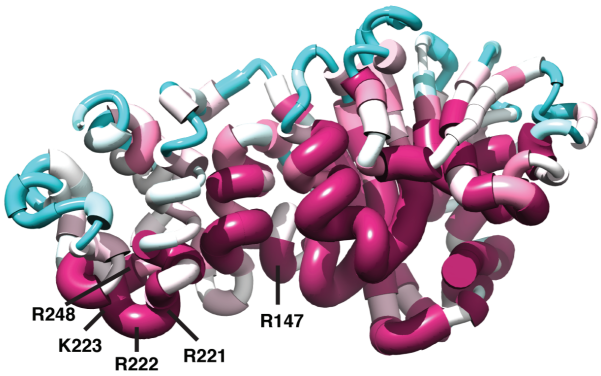

Figure S4.4 Worm representation of KsgA showing the conservation score of each residue. The conservation score is higher for thicker and purple residues and lower for thinner and blue residues. Conservation scores were determined using the ConSurf web server [133].

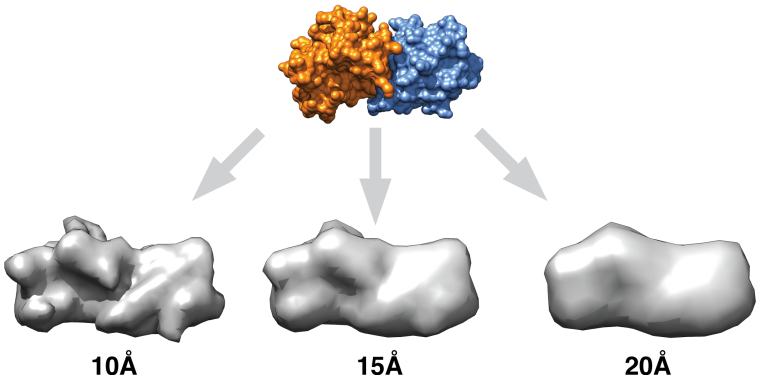

Figure S4.5 Example of simulated cryo-EM data generated for benchmarking HADDOCK-EM. A surface representation of the 7CEI complex is shown on top. Under it, three iso-surfaces are shown for simulated cryo-EM data at 10, 15 and 20Å resolution.

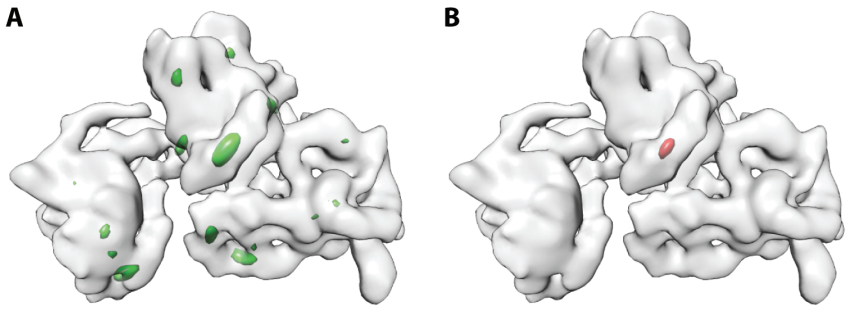

Figure S4.6 Determination of centroid position of KsgA in the cryo-EM density of 30S E.Coli ribosome. Iso-contour of the 30S ribosome in gray and the iso-contour of local cross correlation values at 0.5 (A, green) and 0.6 (B, red) as a result of the full-exhaustive search. The centroid was positioned on the maximum correlation value within the iso-contour shown in B.

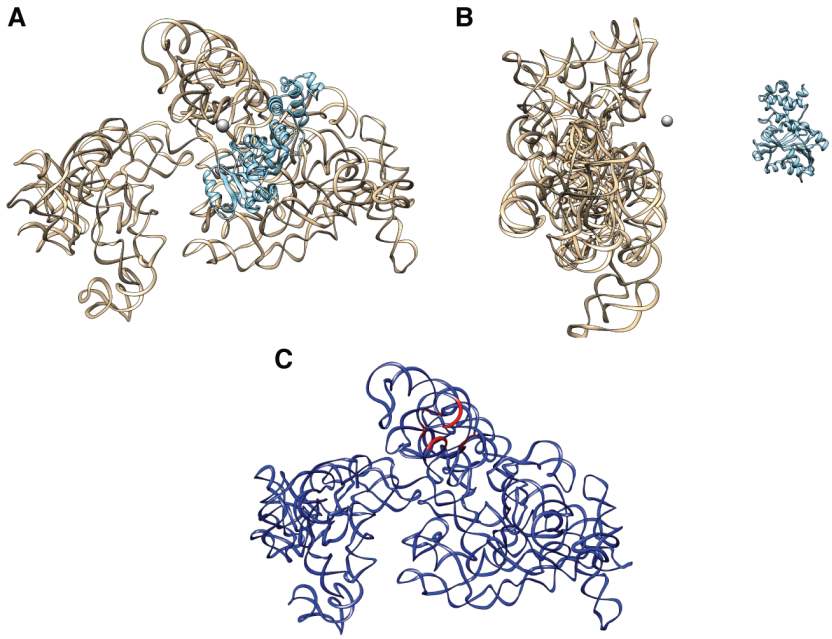

Figure S4.7 Initial placement of the 16S rRNA and KsgA during it0, and active residues of the 16S rRNA. A front- (A) and side-view (B) of the 16S rRNA and KsgA initial setup as was used during the rigid body docking stage. (C) A ribbon representation of the 16S rRNA with the active residues 768 – 773, 781, 782, 801 – 803, 899 – 902, 1512 – 1516 and 1523 shown in red.

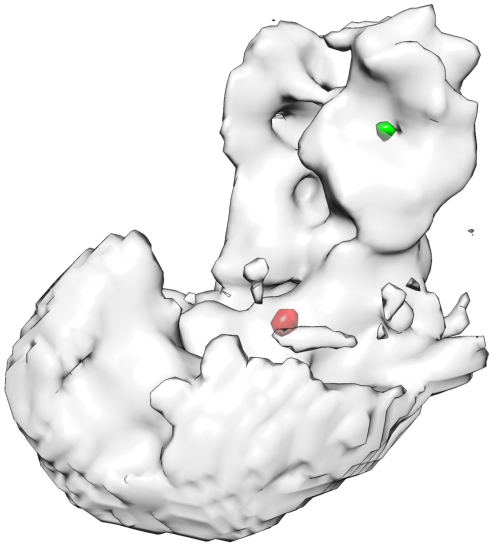

Figure S4.8 Determination of centroid positions for the Dengue-virus envelope protein and antibody. Iso-contour of a subunit part of the 21Å resolution cryo-EM data of Dengue virus (grey), showing regions of high local cross correlation values (0.35) for the envelope protein (red) and antibody (green).

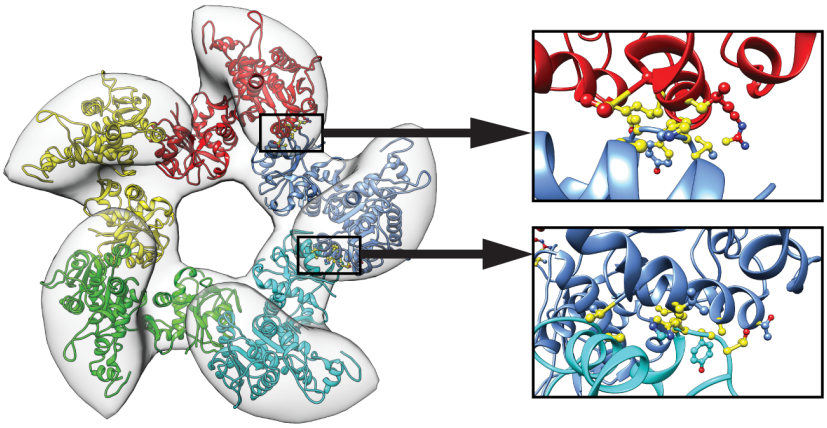

Figure S4.9 Current deposited model of the large terminase complex. A ribbon representation of the current deposited terminase complex (4BIJ). Multiple clashes (yellow ball-and-sticks) are observed when zooming in on the interfaces of the subunits.

Table S4.1 Number of acceptable solutions in the top 400 after each docking stage, using simulated cryo-EM data at 10, 15 and 20Å resolution.

PDB	10Å			15Å			20Å		
	It0	It1	Itw	It0	It1	Itw	It0	It1	Itw
1ACB	13	19	19	7	8	7	4	4	4
1AHW	26	42	42	13	17	17	5	8	8
1AK4	78	81	81	37	40	40	3	8	8
1AVX	38	51	53	13	17	18	5	6	6
1AY7	23	28	28	10	16	16	9	9	9
1B6C	40	47	47	21	28	28	7	12	12
1BGX	0	0	0	0	0	0	0	0	0
1BKD	0	5	5	0	1	1	0	0	0
1JMO	0	0	0	0	0	0	0	0	0
1M10	0	2	2	0	1	1	0	1	1
1R6Q	54	57	57	12	30	30	3	4	4
1JK9	118	223	224	64	84	95	21	23	23
2FD6	140	147	148	82	97	97	37	42	42
2OOB	15	15	15	3	3	4	0	0	0
2OUL	49	62	64	28	32	32	9	10	10
4CPA	90	105	108	85	88	91	65	70	72
7CEI	75	90	92	27	48	49	13	19	20

Table S4.2 Number of acceptable solutions generated with displaced centroids after itw using simulated cryo-EM data at 10, 15 and 20Å resolution.

PDB	10Å				15Å				20Å			
	0Å	3Å	5Å	7Å	0Å	3Å	5Å	7Å	0Å	3Å	5Å	7Å
1ACB	19	12	26	11	7	6	22	4	4	8	1	3
1AHW	42	44	30	32	17	21	20	22	8	10	10	9
1AVX	53	48	57	36	18	23	28	19	6	11	15	16
1JK9	224	198	189	191	95	78	77	28	23	26	10	33
1M10	2	0	3	0	1	1	0	1	1	1	0	0

Table S4.3 Number of acceptable solutions in top 400, 10 and 1 after water refinement stage using ab initio HADDOCK (HADDOCK-CM), and HADDOCK-EM using simulated 20, 15 and 10Å resolution cryo-EM data.

PDB	HADDOCK-CM				HADDOCK-EM 20Å				HADDOCK-EM 15Å				HADDOCK-EM 10Å			
	Total	Top 400	Top 10	Top 1	Total	Top 400	Top 10	Top 1	Total	Top 400	Top 10	Top 1	Total	Top 400	Top 10	Top 1
1AVX	0	0	0	0	6	6	5	1	18	18	10	1	53	53	10	1
2OUL	2	2	1	1	10	10	9	1	32	32	10	1	64	64	10	1
1AY7	1	1	0	0	9	9	8	1	16	16	10	1	28	28	10	1
4CPA	1	1	0	0	72	72	8	1	91	91	10	1	108	108	10	1
1AHW	0	0	0	0	8	8	5	1	17	17	10	1	42	42	10	1
7CEI	4	4	1	0	20	20	5	1	49	49	10	1	92	92	10	1
2OOB	1	1	0	0	0	0	0	0	4	4	0	0	15	15	0	0
2FD6	2	2	0	0	42	42	10	1	97	97	10	1	148	148	10	1
1AK4	1	1	0	0	8	8	3	0	40	40	10	1	81	81	10	1
1B6C	0	0	0	0	12	12	6	1	28	28	10	1	47	47	10	1
1BGX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1R6Q	0	0	0	0	4	4	2	1	30	30	10	1	67	67	10	1
1M10	0	0	0	0	1	1	1	0	1	1	1	1	2	2	2	1
1ACB	0	0	0	0	4	4	1	0	7	7	0	0	19	19	8	1
1JK9	0	0	0	0	23	23	7	1	95	95	10	1	224	224	10	1
1BKD	0	0	0	0	0	0	0	0	1	1	1	0	5	5	5	1
1JMO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4

Chapter 5

The HADDOCK2.2 web server: User-friendly integrative modeling of biomolecular complexes

5.1 Introduction

Cellular metabolism is a highly regulated and adaptive system where proteins, the main participants, form a vast network of interactions collectively known as the interactome. Knowledge of the three dimensional (3D) atomic structure of protein-protein interactions is therefore critical for a fundamental understanding of cellular and molecular biology, as well as for rational drug-design. Unfortunately, solving such structures using classical high-resolution methods (X-ray crystallography and NMR spectroscopy) is not trivial, as each has its own limitations (e.g. protein flexibility, size, strength of the interaction). Considering the magnitude of the interactome, complementary high-throughput methods such as computational docking are necessary if we aim to close the structure gap [104]. The goal of protein-protein docking is to predict the structure of a complex starting from the individual structures of its components [14], which can either be experimentally determined or predicted [134].

Despite continuous advances in the field, the accuracy of ab initio docking – without using any experimental restraints – remains generally low [12]. Data-driven approaches such as HADDOCK [42, 43], which integrate information derived from biochemical, biophysical or bioinformatics methods to enhance sampling, scoring, or both [14], perform remarkably better. The information that can be integrated is quite diverse:

interface restraints from NMR, mutagenesis experiments, or bioinformatics predictions [24, 135]; shape data from small-angle X-ray scattering [64] and cryo-electron microscopy experiments (**Chapter 4**); and orientations of the individual structures in the complex from NMR residual dipolar couplings [57], relaxation anisotropy [58] and pseudocontact shifts experiments [63]. The potential of data-driven docking is reflected in the success of the HADDOCK server and software in recent CAPRI experiments (Critical Assessment of Protein Interaction) [136, 137], as well as in the number of structures deposited (>120) in the Protein Data Bank (PDB), which were calculated using our software.

Five years ago, we introduced the HADDOCK web server to provide a user-friendly interface to the software and streamline its usage by non-expert users in the structural biology field [66]. Shortly after, it was updated to handle multi-body docking [62]. The development of new and improved protocols and the inclusion of additional sources of restraints culminated in the recently released version 2.2 of the software, followed by an update of the web server interfaces. Throughout the next section, we will provide an overview of the newly updated HADDOCK web server and discuss the most relevant additions. The server is freely accessible at **http://haddock.science.uu.nl/services/HADDOCK2.2** to non-profit users upon registration. We conclude by presenting usage statistics of the server to demonstrate the usefulness and power of providing easy and free access to scientific software.

5.2 Overview and advances

The HADDOCK web server was created to facilitate the use of our docking software, by removing the burden of its installation and setup, as well as by providing validation routines for input data and options. In addition, since HADDOCK runs are computationally demanding, the web server offers the users access to sufficient resources - our local cluster(s) - to complete their runs within a few hours. A grid-enabled version of the server can be accessed via the WeNMR web site (**http://www.wenmr.eu**) [138], which uses resources provided by the European Grid Initiative (EGI) and the associated National Grid Initiatives (NGIs). This setup, which currently handles most submissions, provides more than 110.000 CPU cores distributed over 41 sites worldwide (see **http://gstat.egi.eu/gstat/geo/openlayers#/VO/enmr.eu**).

The HADDOCK web server aggregates seven different interfaces, each associated with a different level of control over the docking protocol reflected by the number of parameters that can be changed. New users are granted access to the Easy and the associated Prediction Interface only, but can request access to the Expert and Guru levels, and their associated interfaces, if necessary.

The Easy interface provides the most basic level of control. It allows the user to either upload two structures in PDB format or download them directly from the RCSB PDB, and define sets of active and passive residues that represent the (putative) interface. Unlike previous versions, HADDOCK 2.2 supports single (protein, small molecule, RNA, or DNA) and mixed (protein-DNA, protein-RNA) molecule types. This was implemented to handle the docking of proteins onto a nucleosome complex – a recent CAPRI target.

The Expert interface builds on the Easy interface and allows the user to manually specify the protonation state of each histidine residue in the proteins, which is otherwise determined automatically with MolProbity [139]. Also, it offers control over which regions of the molecules are semi-flexible and fully flexible segments, which has an impact during the refinement stage of the docking. Lastly, the user is given the option to define the charge state of the N- and C-terminus of the protein. The Expert interface also provides a Distance Restraints section, where the user has the option to upload user-defined ambiguous and unambiguous restraints files and/or use center-of-mass restraints, useful for blind or ab initio docking when no other information is available, but also to ensure compactness of the generated models. The center-of-mass restraints are automatically generated by calculating the dimensions of each molecule along the x, y and z-axis (d_x, d_y, d_z) and summing the average of the two smallest components per molecule. The resulting distance is used to define a restraint between the center of mass of each subunit with an additional upper bound corrections of 1Å [64]. In addition, the Expert interface gives control over the Sampling Parameters, including namely the number of structures to generate at each stage and whether or not to perform solvated docking [60, 140, 141]. Finally, it exposes the Clustering Parameters that define the clustering algorithm and cutoff. In version 2.2, in addition to RMSD-based clustering, there is the option of using the Fraction of Common Contacts (FCC) clustering algorithm [142], which is significantly faster and especially useful for symmetric complexes.

The Guru interface gives full access and control to ~500 parameters, nearly all that are available in HADDOCK. The Distance Restraints section now offers a new radius of gyration restraint, information that can be extracted, for example, from SAXS experiments. Non-crystallographic Symmetry Restraints and Symmetry Restraints are also available at this level and have been extended to handle C4- and D2-symmetries in addition to the already available C2-, C3- and C5-symmetries. There are additional sections for other types of NMR-based restraints, such as Residual Dipolar Couplings [57], Relaxation Anisotropy [58], and the recently added Pseudo Contact Shifts [63]. These latter require a tensor distance restraints file and the definition of the rhombic and axial components of the anisotropic tensor. Besides the restraints, all the energy evaluations, scoring functions and analysis parameters can be tweaked; advanced parameters for the sampling protocols are also available at this level, offering a greater degree of control, for example, on the extent of each refinement stage. There are also dedicated options to the solvated docking protocol, which now uses by default propensities based on the Kyte-Doolittle hydrophobicity scale, as these have been shown to improve the protocol [140]. The original statistical-based propensities [60], recently expanded to include nucleotides [141], can still be selected via a dropdown menu.

The remaining four interfaces consist of: the Prediction Interface, which is similar to the Easy interface, but with settings geared towards using bioinformatics interface predictors such as CPORT [135]; the Refinement Interface (expert-level access), which runs only the water refinement stage on the uploaded structures and can be used for scoring purposes; the Multi-body Interface, based on the Guru interface, supporting upload of up to six molecules that will be docked simultaneously [62] and also featuring the Molecule Interaction Matrix section. This new addition displays a table with scaling factors to adjust the interaction forces between different subunits, allowing molecules to become invisible to each other during the docking, which is useful in cases where multiple binding modes are required to satisfy the experiment data (see for example [143]). The use of ambiguous interaction restraints within this interface requires the user to upload a restraints table file in the Distance Restraints section This requirement of uploading distance restraint files instead of supplying residue lists as in the other interfaces was meant to

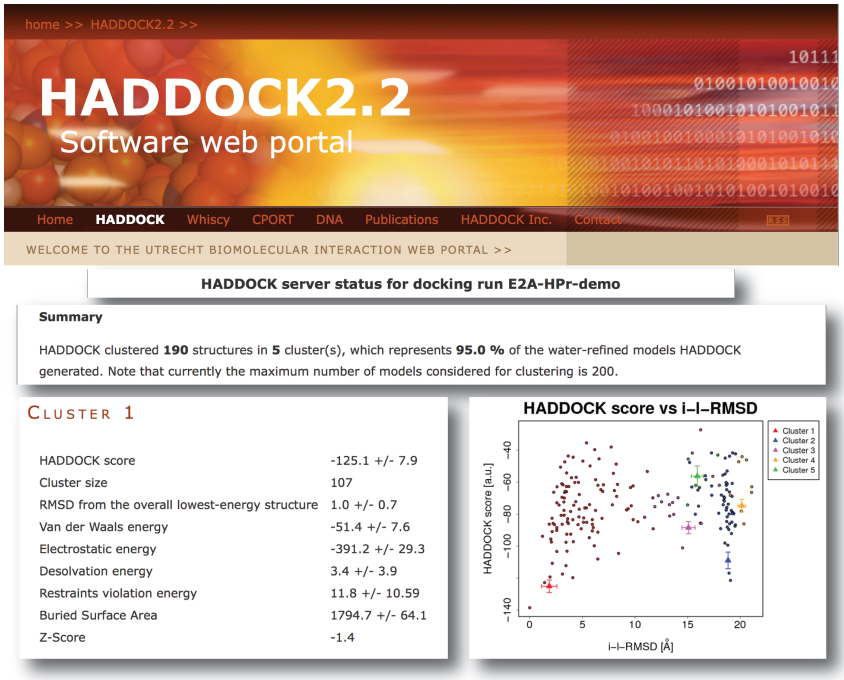

Figure 5.1 Excerpts from an example result page of a HADDOCK2.2 docking run.

make users think carefully about their system since in multi-body docking multiple interfaces will be defined that might not all be supposed to interact. To facilitate the creation of custom ambiguous interaction restraint files between any number of molecules, an interface called Gentbl was created. Finally, the web server also offers a File Upload Interface to allow the user to upload a run parameter file, created upon successful validation and submission to the queue, and thus easily redo a docking run or re-run it with slight changes in the parameters.

At submission time, once the input data have been properly validated, the server offers the option to download a parameter file, and provides a link to the results page, which is also emailed to the user. Users are encouraged to save the parameter file since it contains all required input data and settings to reproduce the docking, as recommended in the "Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop" (Recommendation 1) [144]. The results page allows monitoring of the progress of the docking run. After a successful docking run, the user will receive another e-mail redirecting him/her to the updated

results page (see **Figure 5.1** for an excerpt of presented results). The page indicates how many structures of the water-refined models could be clustered, and lists the clusters in the order of their HADDOCK score. For each cluster, detailed statistics are displayed, representing the average values calculated over the top four best scoring structures within each cluster. Besides the HADDOCK score and other standard energies (van der Waals, etc.), a z-score has been added. The z-score represents how many standard deviations the HADDOCK score of a given cluster is separated from the mean of all clusters, i.e. the lower the z-score, the better. To visualize the results, plots are displayed at the bottom of the results-page, showing for example the HADDOCK score of all solutions against the interface-ligand-RMSD (i-l-RMSD) compared to the best scoring structure, together with cluster averages and their spreads.

5.3 Usage statistics

Since its opening in June 2008, the HADDOCK web server has seen a sustained increase in the number of registrations to reach over 6000 registered users to date distributed all over the world (**Figure 5.2**). More than 108,000 runs have been processed, 28% of which have run on EGI grid resources. This percentage has increased to 75% for the HADDOCK2.2 server submission. An overview of the number of runs processed per month with their distribution over local and grid resources is shown in **Figure 5.2**. Since the launch of the HADDOCK2.2 web server in March 2015, an increased fraction of runs are handled by the new 2.2 portal. Statistics over the last year (since January 2014) indicate that the portal is processing on average 75 docking runs per day. The execution wall time averaged over both local cluster and grid resources is around 16 hours per HADDOCK run, as approximately 75 runs are handled per day with up to 50 jobs running in parallel (and a maximum of 5 concurrent jobs per user). The exact run time depends on the size of the system being docked and the parameter settings and can vary between half an hour and several days. The server home page reports both the number of running and pending jobs, allowing users to get an estimate of the waiting time. A majority of the docking runs are dealing with protein-protein and protein-peptide docking (~61%), ~19% correspond to protein-nucleic acids systems and quite a significant fraction

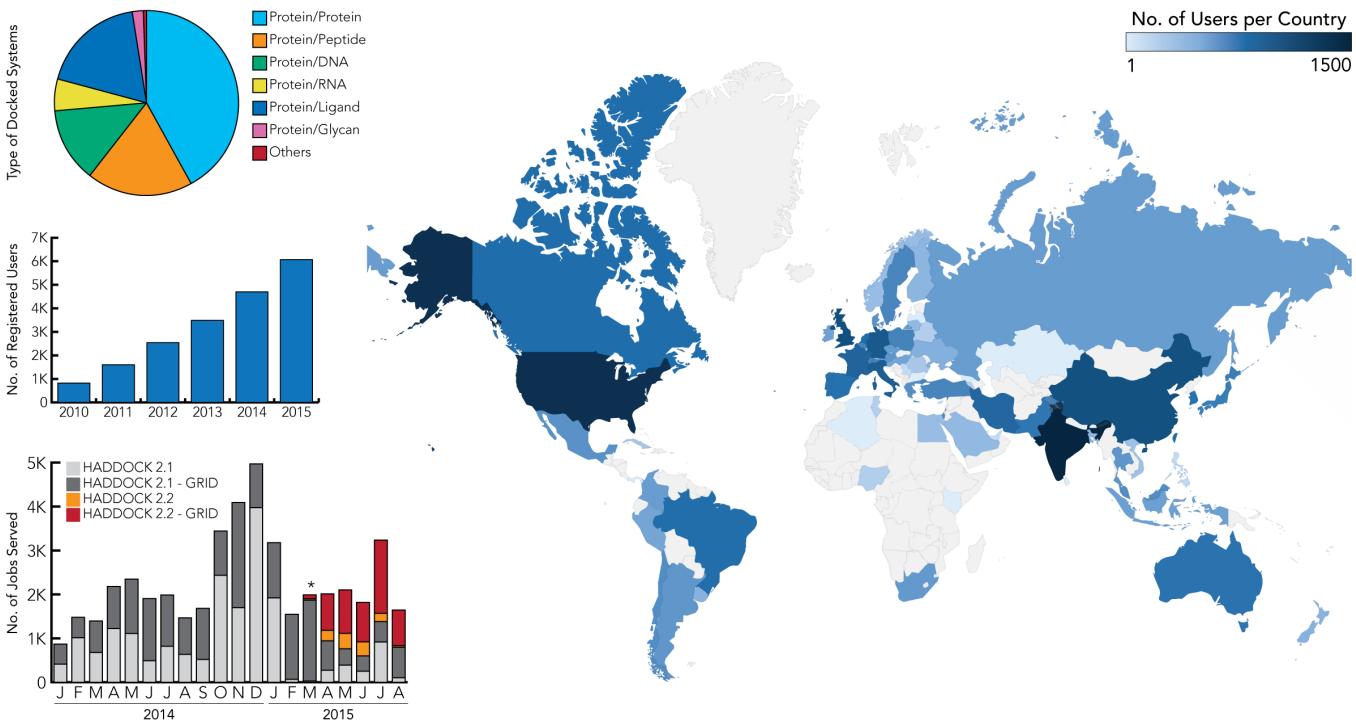

Figure 5.2 HADDOCK web server statistics (as of September 1st 2015). The worldwide distribution of registered users (>6000) is shown with a blue-scale color-coding. On the top right, the distribution of the type of systems docked using HADDOCK is shown (taken from a HADDOCK online survey with >325 respondents). The middle bar graph shows the yearly increase in the number of registered users since 2010 and the bottom plot shows the number of processed user runs per month since last year (2014), distinguishing between runs that have been processed on our local resources or on the EGI grid resources, and between the previous and current (2.2) version of the server.

(~20%) is dealing with protein-small molecule docking (both small ligand and oligosaccharides). These numbers demonstrate the popularity and widespread usage (both in terms of geographic distribution and type of systems being studied) of our HADDOCK web server.

Since its opening in June 2008, the HADDOCK web server has seen a sustained increase in the number of registrations to reach over 5500 registered users to date distributed all over the world (**Figure 5.2**). More than 103,000 runs have been processed, 28% of which have run on EGI grid resources. This percentage has increased to 75% for the HADDOCK2.2 server submission. An overview of the number of runs processed per month with their distribution over local and grid resources is shown in **Figure 5.2**. Since the launch of the HADDOCK2.2 web server in March 2015, an increased fraction of runs are handled by the new 2.2 portal. Statistics over the last year (since January 2014) indicate that the portal is processing on average 75 docking runs per day. A majority of these runs are dealing with protein-protein and protein-peptide docking (~61%), ~19% correspond to protein-nucleic acids systems and quite a significant fraction (~20%) is dealing with protein-small molecule docking (both small ligand and oligosaccharides). These numbers demonstrate the popularity and widespread usage (both in terms of geographic distribution and type of systems being studied) of our HADDOCK web server.

Notes

This Chapter is based on: G.C.P. van Zundert, J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, E. Karaca, A.S.J. Melquiond, M. van Dijk, S.J. de Vries and A.M.J.J. Bonvin. The HADDOCK2.2 web server: User-friendly integrative modeling of biomolecular complexes. *J Mol Biol*, Advanced Online Publication (2015).

Chapter 6

DisVis: quantifying and visualizing accessible interaction space of distance-restrained biomolecular complexes

6.1 Introduction

Structural characterization of protein complexes is of paramount importance for a fundamental understanding of cellular processes, and with major applications in rational drug-design. As the quantity of experimentally determined complexes is only a fraction of their total predicted number, complementary computational techniques have been developed for predicting the structure of complexes from their components [14, 104]. Additional low-resolution information in the form of distance restraints can significantly benefit the modeling, with a variety of experimental methods providing such information, such as chemical cross-links detected by mass spectrometry [145], and distance measurements from electron paramagnetic resonance (EPR) and FRET [146].

When two biomolecules are known to interact and no high-resolution model is available, the structure of the complex can naively be any one state where the molecules are in contact. We define the accessible interaction space of the complex as the set of all these states. If a distance restraint is imposed on the complex, the accessible interaction space reduces, depending on the information content of the restraint. The interaction space is further reduced if multiple restraints are included. So far, however, no computational method has been reported that quantifies this reduction or allows to visualize this accessible interaction space.

To aid in this task, we have developed DisVis, a GPU-accelerated Python software package and command line tool (*disvis*) for quantifying and visualizing the accessible interaction space of distance-restrained binary complexes. Disvis takes as input two atomic structures and a file with distance restraints, and outputs the sum of complexes complying with a given number of restraints together with a density showing the maximum number of consistent restraints at every position in space. This indicates whether all data are consistent and can be combined without violations, and allows identification of false positives, quantification of the information content of the restraints and visualization of interesting regions in the interaction space. The method is generic and can easily be incorporated into existing Fast Fourier Transform (FFT)-accelerated docking programs as a distance-dependent energy function, allowing the 'marriage made in heaven' of direct sampling and scoring of FFT-generated docking poses [147] at a small computational cost.

6.2 Methods

6.2.1 Overview

We discretely sample the accessible interaction space by treating the two biomolecules as rigid bodies and performing a 6 dimensional search over the three translational and three rotational degrees of freedom. We use FFT-techniques to accelerate the translational search using a 1Å grid spacing (default). These have long been used in the docking field [148]. One chain is fixed in space and considered the receptor molecule, while translational scans are performed for each rotation of the ligand molecule. Two atoms i and j are considered to be interacting if the distance, d, between them is $r_{\mathrm{vdW}} < d \leq r_{\mathrm{vdW}} + 3$Å (by default), where r_{vdW} is the combined van der Waals radius of the two atoms $r^i_{\mathrm{vdW}} + r^j_{\mathrm{vdW}}$, and clashing if $d \leq r_{\mathrm{vdW}}$. A conformation is deemed a complex if the volume of interaction is above- and the volume of clashes below threshold values (300 and 200Å^3 by default, respectively).

After every translational scan, all conformations that comply with each restraint are determined. Next, *disvis* counts the number of accessible complexes consistent with each number of restraints, as well as which restraints are violated. This is repeated until the rotational sampling

density reaches a pre-set value (default 9.72°, 7416 orientations). During the rotational search, *disvis* stores the maximum number of consistent restraints found at every scanned position of the ligand's center of mass, which ultimately results in a discrete 'density' map. The output thus consists of the sum of accessible interaction statescomplexes complying with each number of restraints, a percentage of how often each restraint is violated, and a discrete-valued density map.

6.2.2 Calculating the accessible interaction space of two interaction macromolecules

As a first approximation to calculate the accessible interaction space of two interacting macromolecules and to make the computation more tractable, we treat the molecules as rigid entities. This results in a 6 dimensional (3 translational and 3 rotational degrees of freedom) space of possible conformations that need to be considered. To determine within this 6D space whether the two chains are interacting and forming a complex, we use FFT-techniques as used originally in Katchalski-Katzir et al. [148]. We keep one chain fixed during the search while we perform FFT-accelerated translational scans with the other chain. The fixed chain is separated into a core and interaction region. The core region is the space that is occupied by combining spheres with each center at the atom coordinate and as radius the elements' van der Waals radius; the interaction region is determined similarly, but the radius is extended by 3Å (by default). The 3D shapes are subsequently projected onto a grid with a voxel spacing of 1Å (by default). The scanning chain is only represented by its core object. The resulting shape is again projected onto a grid with equal voxel spacing as the fixed chain to allow for FFT-accelerated translational scans during the search.

After the creation of the search objects, we identify clashes and interactions as a function of rotation R as follows

$$\mathbf{C}(R) = \mathcal{F}^{-1}\left[\mathcal{F}(\mathbf{S}(R))^* \times \mathcal{F}(\mathbf{F}_{\text{core}})\right] \tag{6.1}$$

$$\mathbf{I}(R) = \mathcal{F}^{-1}\left[\mathcal{F}(\mathbf{S}(R))^* \times \mathcal{F}(\mathbf{F}_{\text{inter}})\right] \tag{6.2}$$

where the cross-correlation theorem has been used to calculate $\mathbf{C}$ and $\mathbf{I}$, the spaces that represent the volume of clashes and interactions at every

grid position in Å^3, respectively; $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the FFT operator and its inverse, respectively; * is the complex conjugate operator, and $\times$ the elementwise multiplication operator; $\mathbf{S}$ is the shape of the scanning chain, and $\mathbf{F}_{\text{core}}$ and $\mathbf{F}_{\text{inter}}$ are the core and interaction shapes of the fixed chain, respectively.

To determine whether a conformation is a plausible complex its clashing volume should not be too large, while the interaction volume should be of reasonable size. The accessible interaction space per translational space is then given by

$$\mathbf{A}_R(\vec{r}) = \begin{cases} 1 & \text{if } \mathbf{C}_R(\vec{r}) \leq \mathbf{C}_{\max} \text{ and } \mathbf{I}_R(\vec{r}) \geq \mathbf{I}_{\min} \\ 0 & \text{else} \end{cases}$$

where $\mathbf{C}_{\max}$ and $\mathbf{I}_{\min}$ are parameters representing the allowed maximum volume of clashes (200Å^3 by default) and the minimum volume of interactions (300Å^3 by default), respectively. Raising $\mathbf{C}_{\max}$ and lowering $\mathbf{I}_{\min}$ results in a more lenient counting of accessible states, while lowering $\mathbf{C}_{\max}$ and raising $\mathbf{I}_{\min}$ makes the counting for accessible states more stringent. The total number of accessible states is determined by performing an exhaustive search over rotation space and counting at every rotation all states where $\mathbf{A}_R$ equals 1. Care should be taken here that rotation space is as evenly and optimally sampled as possible to minimize redundancy and biasing certain orientations in the counting. To take this into account, we used the optimal rotation sets developed by Karney [80], which include a weight factor for every rotation to average out redundancy. The total number of accessible states N_A is thus given by

$$N_A = \sum_{\mathbf{P}} w_R \sum_{x,y,z} \mathbf{A}_R(x,y,z) \tag{6.3}$$

where w_R is the weight factor for the specific orientation/rotation, the first summation is over all rotations $\mathbf{P}$, and the second summation over all grid coordinates.

6.2.3 Incorporating distance restraints into the search

If some distances or distance ranges are known between the subunits of the complex, this can significantly reduce the accessible interaction

space as it puts extra restraints on the requirement for a conformation to be considered a complex. To combine this information with FFT-accelerated translational scans, the whole space of conformations that comply with the distance restraint should be demarcated for every rotation at once. As the distance of the restraint depends only on the coordinates of two atoms (or points, more generally), the space consistent with the restraint must be represented by a sphere with a radius corresponding to the distance restraint. The remaining parameter that needs to be determined is the position of the center of this sphere $\vec{r}_c$, which is given by

$$\vec{r}_c\,(R) = \vec{r}_{\mathrm{F}} - [\vec{r}_{\mathrm{S}}\,(R) - \vec{r}_{\mathrm{comS}}\,(R)] \tag{6.4}$$

where $\vec{r}_{\mathrm{F}}$ and $\vec{r}_{\mathrm{S}}$ are the coordinates of the restrained atoms of the fixed and scanning chain, respectively, and $\vec{r}_{\mathrm{comS}}$ is the center of mass of the scanning chain. The equation can be simplified by initially placing the center of mass of the scanning chain on the origin, and rotating the scanning chain around its center of mass. Furthermore, realizing that $\vec{r}_{\mathrm{F}}$ is fixed and $\vec{r}_{\mathrm{S}}$ now only depends on the rotation of the scanning chain, **Eq. 6.4** reduces to

$$\vec{r}_c\,(R) = \vec{r}_{\mathrm{F}} - R\vec{r}_{\mathrm{S}} \tag{6.5}$$

The space of states complying with the distance restraint per translational scan $\mathbf{L}_R$ is defined then as

$$\mathbf{L}_R\,(\vec{r}) = \begin{cases} 1 & \text{if } d_{\min} \leq |\vec{r} - \vec{r}_c| \leq d_{\max} \\ \\ 0 & \text{else} \end{cases}$$

where $d_{\min}$ and $d_{\max}$ are the minimum and maximum allowed distance, respectively. Note that the function describing $\mathbf{L}_R$ can be freely chosen, under the restriction that it should be spherical symmetric, which opens up the use of more complex distance restraints in FFT-docking software. The reduced accessible interaction space $\mathbf{A}_{R,\mathrm{red.}}$ is then simply given by

$$\mathbf{A}_{R,\mathrm{red.}} = \mathbf{A}_R \times \mathbf{L}_R \tag{6.6}$$

In the case of multiple available distance restraints, this generalizes to

$$\mathbf{A}_{R,\text{red.}} = \mathbf{A}_R \times \sum_{n}^{N_d} \mathbf{L}_{R,n} \tag{6.7}$$

where the summation is over all distance restraints N_d and $\mathbf{L}_{R,n}$ is the space conforming to distant restraint n. The value found at a specific coordinate in $\mathbf{A}_{R,\text{red.}}$ represents the number of conforming distance restraints at that location in space.

6.2.4 Quantifying and visualizing the accessible interaction space

To quantify the accessible interaction space consistent with a certain number of distance restraints, the number of occurrences that $\mathbf{A}_{R,\text{red.}}$ is equal to the number of compliant restraints is counted. The accessible interaction space is visualized by outputting the maximum value found during the rotational search at every grid position, thus given by

$$\mathbf{V}(x, y, z) = \max\left\{\mathbf{A}_{R,\text{red.}}(x, y, z) : R \in \mathbf{P}\right\} \tag{6.8}$$

The resulting 'density' is written to file in MRC format and represents the position of the center of mass of the scanning chain relative to the fixed chain. These files can straightforwardly opened with molecular visualization programs, such as PyMol and UCSF Chimera. With this information, interesting regions of high- density can then be sampled more thoroughly. Also, false-positive restraints can be identified if the exhaustive search does not result in a region where all restraints of the cross-links are obeyed. In addition, for each complex that is consistent with at least one restraint, all restraints that are violated are calculated and stored during the search. This ultimately results in a violation matrix where every row represents the number of consistent restraints and every column indicates how often a specific restraint is violated for complexes consistent with at least N restraints (e.g. **Table 6.3**). Lastly, to give the user an indication which restraints are most likely to be false-positives, the z-score is calculated for each restraint based on the violation matrix given by

$$Z = \frac{v_i - \bar{v}}{\sigma} \tag{6.9}$$

where v_i is the column average of the violation matrix of restraint i, and $\bar{v}$ and σ are the average and standard deviation of the violation matrix. *Disvis* reports restraints with a z-score higher than 1.0 explicitly.

6.2.5 Implementation details

We implemented DisVis in Python2.7 using the NumPy [149] and Cython packages [150]. The OpenCL framework [151] was used to offload the computations to the GPU. Python bindings were available through the pyopencl package [152]. We used the high-performance clFFT library (`https://github.com/clMathLibraries/clFFT`) together with gpyfft for Python bindings (`https://github.com/geggo/gpyfft`) to calculate the FFTs. Computations were performed on AMD Opteron 6344 CPU processors and on an AMD Radeon HD 7730M and NVIDIA GeForce GTX 680 GPU. DisVis code can be downloaded freely from `https://github.com/haddocking/disvis` together with documentation and examples.

6.2.6 RNA polymerase II example

The crystal structure of the RNA polymerase II was downloaded from the Protein Databank (PDB ID: 1WCM). The largest subunit (chain A) and the 27kDa polypeptide (chain E) were extracted from the PDB-file. Six BS3 cross-links were available and taken from XLdb [47]. To investigate the detection of false-positive restraints, two virtual cross-links were added with a distance of 35.7 and 42.2Å using the Xwalk web server [153]. The maximum allowed distance of the BS3 cross-links was set to 30Å, based on molecular dynamics trajectory analysis [154]. The restraints used are shown in **Table S6.1**. The input files are included in the DisVis source code.

Two *disvis* runs were performed using a rotational sampling density of 5.27° (53256 orientations) and 9.72° (7416 orientations) with a grid spacing of 1 and 2Å, respectively. All parameters were left to their default values (interaction radius 3Å, minimum required volume of interaction 300Å^3 and maximum allowed volume of clashes 200Å^3). The number of accessible complexes consistent with each number of cross-links is shown in **Table S6.2** and **S6.4**, and the relative occurrence of

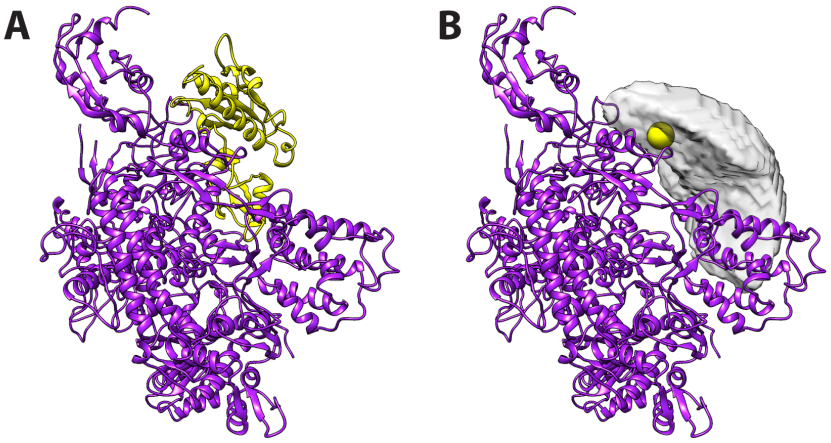

Figure 6.1 Accessible interaction space of two chains of the RNA polymerase II complex. (A) The large subunit (purple) and the 27kDa polypeptide (yellow) . (B) The large subunit and the reduced accessible interaction space of the 27kDa polypeptide consistent with at least 6 cross-links in grey. The smooth yellow sphere represents the center of mass of the polypeptide.

restraint violations in **Table S6.3** and **S6.5** for the fine and coarse run, respectively.

6.2.7 26S proteasome PRE5-PUP2 example

Homology models were downloaded from the SWIS-MODEL Repository [155] via the Protein Model Portal (`http://proteinmodelportal.org`) using their Uniprot identifiers (O14250 and Q9UT97). Cross-links were taken from Leitner et al. [156] Dataset S1 (**Table S6.6**), which consist of 4 ADH and 3 zero-length ZL cross-links. The maximum ADH- and ZL-linker length were set to 23 and 26Å, respectively, since 95% of all found distances in a benchmark were smaller. All input files are included in the DisVis source code.

Again, two *disvis* runs were performed using a rotational sampling density of 5.27° (53256 orientations) and 9.72° (7416 orientations) with a grid spacing of 1 and 2Å, respectively, with default parameter values. The sum of accessible complexes consistent with each number of restraints is shown in **Table S6.7** and **S6.9**, and their normalized restraint violation occurrence in **Table S6.8** and **S6.10**.

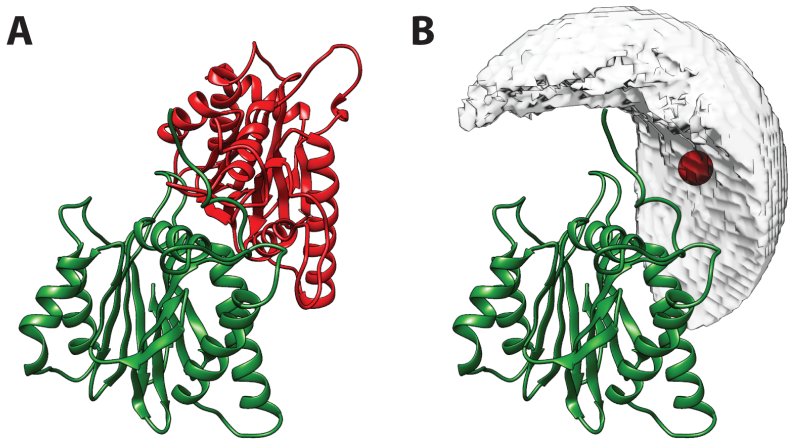

Figure 6.2 Accessible interaction space of the PRE5-PUP2 complex. (A) The PRE5 (red) complexed with PUP2 (green), based on a homology model. (B) PRE5 and the accessible interaction space of PUP2 consistent with all 7 distance restraints (grey). The smooth red sphere represents the center of mass of PUP2.

6.3 Examples

To illustrate the capabilities of *disvis*, we applied it on two systems, using MS cross-links data. A fine rotational search (5.27°, 53256 orientations) was performed using default values. First we investigated the accessible interaction space of two chains of the RNA polymerase II complex of S. cerevisiae (1WCM, chain A and E) for which 6 BS3 cross-links were available (**Table S6.1**) [47, 157]. The allowed distance was set between 0 and 30Å (C_β– C_β) for every restraint. Two false-positive restraints were added with a distance in the crystal structure of 35.7 (FP1) and 42.2Å (FP2) to test whether these violating restraintsthey could be identified. Applying disvis shows that none of the 18.9×10^9 complexes sampled are consistent with all 8 restraints, though a small number are conforming to 7 cross-links (9716 complexes) (**Table S6.2**). For the latter, only restraint FP2 is violated. The accessible interaction space consistent with at least 6 restraints is less than 0.03% of the full interaction space (**Figure 6.1**). The density clearly indicates the position of the E-chain. Interestingly, both false-positive restraints are violated in 100% of the complexes consistent with at least 6 restraints; in contrast, the highest violation percentage of a correct cross-link is only 0.1% (**Table S6.3**). Thus, a high-violation percentage is an indication of a false-positive restraint.

Secondly, we applied disvis on two proteins of the 26S proteasome of S. pombe, PRE5 and PUP2, with 7 cross-links available (**Table S6.6**) [156]. The acceptable distances for the ADH and ZL cross-links were set to 23 and 26Å (C_α– C_α), respectively, as 95% of distances found in a benchmark were shorter [156]. The PRE5-PUP2 complex is significantly smaller than the previous example with the full interaction space consisting of 6.9×10^9 complexes. Still, the accessible interaction space consistent with all 7 restraints is heavily reduced to less than 0.04% of the full interaction space. The accessible interaction space of the PUP2 chain with respect to PRE5 is overlapping with its center of mass deduced from a homology model (**Figure 6.2**).

The computation for those two examples took 74m and 27m on 16 AMD Opteron 6344 processors and 76m and 19m on an NVIDIA GeForce GTX 680 GPU, respectively. However, by increasing the voxel spacing to 2Å and using a coarser rotational search (9.72°, 7416 orientations) rather similar results can be obtained in only 19m and 8m, respectively, on a single processor (cf. **Table S6.2** and **S6.4** for example). It should further be noted that the bulk of the time is spent on computing the FFTs and a negligible part on computing the consistent distance restraint space (**Table S6.11**).

6.4 Conclusions

We have introduced DisVis, a Python package and command line tool to quantify and visualize the information content of distance restraints, and a powerful aid in detecting the presence of false-positive restraints. Our novel approach can be easily incorporated in FFT-accelerated docking programs, allowing the use of any form of distance-dependent energy function.

Notes

This Chapter is based on: G.C.P. van Zundert and A.M.J.J. Bonvin. DisVis: quantifying and visualizing accessible interaction space of distance-restrained biomolecular complexes. *Bioinformatics*, Advanced Online Publication (2015).

Supplementary information

Table S6.1 Cross-links used to assess accessible interaction space of the large subunit and the 27kDa polypeptide of the RNA polymerase II complex.

Residue chain A	Residue chain E	Cross-linker[a]	Distance in complex (Å)[b]
1003	166	BS3	12.5
129	161	BS3	19.8
129	171	BS3	12.9
15	171	BS3	19.6
934	201	BS3	21.8
938	201	BS3	15.1
180	122	Virtual	35.7
1092	152	Virtual	42.2

[a] Cross-link chemistry. BS3: Bissulfosuccinimidyl suberate; Virtual: Manually added false-positive cross-link.

[b] C_β – C_β distance in crystal structure (1WCM)

Table S6.2 Total number of accessible complex conformations per number of complying restraints of the large subunit and the 27kDa polypeptide of the RNA polymerase II complex using a fine rotational search (5.27°, 53256 orientations) and grid (1Å).

Number of consistent restraints (N)	Number of accessible complexes consistent with exactly N restraints	Fraction of accessible complexes consistent with exactly N restraints	Number of accessible complexes consistent with at least N restraints	Fraction of accessible complexes consistent with at least N restraints
0	16570457037	0.8749	18940752204	1.0000
1	1392884181	0.0735	2370295166	0.1251
2	678488947	0.0358	977410985	0.0516
3	206270378	0.0109	298922038	0.0158
4	74963882	0.0040	92651659	0.0049
5	12515339	0.0007	17687776	0.0009
6	5162720	0.0003	5172437	0.0003
7	9716	0.0000	9716	0.0000
8	0	0.0000	0	0.0000

Table S6.3 Normalized occurrence of a restraint violation given a number of consistent restraints for the large subunit and the 27kDa polypeptide of the RNA polymerase II complex using a fine rotational search (5.27°, 53256 orientations) and grid (1Å).

Number of consistent restraints (N)	Fraction of complexes consistent with N restraints in which a specific restraint is violated							
	Restraint 1	Restraint 2	Restraint 3	Restraint 4	Restraint 5	Restraint 6	Restraint 7	Restraint 8
1	0.731	0.813	0.781	0.813	0.742	0.780	0.772	0.981
2	0.676	0.617	0.586	0.725	0.504	0.497	0.974	0.996
3	0.308	0.344	0.285	0.434	0.654	0.622	0.970	0.996
4	0.080	0.151	0.057	0.238	0.653	0.607	0.968	1.000
5	0.015	0.140	0.001	0.371	0.180	0.061	0.940	1.000
6	0.000	0.000	0.000	0.000	0.001	0.000	0.997	1.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table S6.4 Total number of accessible complex conformations per number of complying restraints of the large subunit and the 27kDa polypeptide of the RNA polymerase II complex using a coarse rotational search (9.72°, 7416 orientations) and grid (2Å).

Number of consistent restraints (N)	Number of accessible complexes consistent with exactly N restraints	Fraction of accessible complexes consistent with exactly N restraints	Number of accessible complexes consistent with at least N restraints	Fraction of accessible complexes consistent with at least N restraints
0	287850752	0.8757	328691520	1.0000
1	24045812	0.0732	40840780	0.1243
2	11681382	0.0355	16794968	0.0511
3	3538009	0.0108	5113586	0.0156
4	1281164	0.0039	1575577	0.0048
5	208446	0.0006	294412	0.0009
6	85798	0.0003	85966	0.0003
7	167	0.0000	167	0.0000
8	0	0.0000	0	0.0000

Table S6.5 Normalized occurrence of a restraint violation given a number of consistent restraints for the large subunit and the 27kDa polypeptide of the RNA polymerase II complex using a coarse rotational search (9.72°, 7416 orientations) and grid (2Å).

Number of consistent restraints (N)	Fraction of complexes consistent with N restraints in which a specific restraint is violated							
	Restraint 1	Restraint 2	Restraint 3	Restraint 4	Restraint 5	Restraint 6	Restraint 7	Restraint 8
1	0.731	0.812	0.780	0.811	0.743	0.782	0.774	0.981
2	0.679	0.615	0.583	0.724	0.507	0.501	0.974	0.997
3	0.313	0.339	0.282	0.431	0.659	0.628	0.970	0.996
4	0.080	0.146	0.055	0.235	0.660	0.615	0.967	1.000
5	0.015	0.135	0.001	0.373	0.182	0.063	0.937	1.000
6	0.000	0.000	0.000	0.000	0.001	0.000	0.996	1.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table S6.6 Cross-links used to assess the accessible interaction space of PUP2 relative to PRE5. Data were taken from Leitner et al. [156] Dataset S1.

Residue PRE5	Residue PUP2	Cross-linker[a]	Distance in complex (Å)[b]
27	18	ADH	5.9
122	125	ADH	12.1
122	127	ADH	5.7
122	128	ADH	7.8
54	179	ZL	9.1
55	169	ZL	10.8
55	179	ZL	10.9

[a] Cross-link chemistry. ADH: adipic acid dihydrazide; ZL: zero-length
[b] C_α – C_α distance in homology model

Table S6.7 Total number of accessible complex conformations per number of complying restraints of the PRE5-PUP2 complex using a fine rotational search (5.27°, 53256 orientations) and grid (1Å).

Number of consistent restraints (N)	Number of accessible complexes consistent with exactly N restraints	Fraction of accessible complexes consistent with exactly N restraints	Number of accessible complexes consistent with at least N restraints	Fraction of accessible complexes consistent with at least N restraints
0	5431316957	0.7837	6930088505	1.0000
1	565217635	0.0816	1498771547	0.2163
2	226110049	0.0326	933553912	0.1347
3	622583287	0.0898	707443862	0.1021
4	73552113	0.0106	84860574	0.0122
5	4747627	0.0007	11308461	0.0016
6	4069363	0.0006	6560833	0.0009
7	2491469	0.0004	2491469	0.0004

Table S6.8 Normalized occurrence of a restraint violation given a number of consistent restraints for the PRE5-PUP2 complex using a fine rotational search (5.27°, 53256 orientations) and grid (1Å).

Number of consistent restraints (N)	Fraction of complexes consistent with N restraints in which a specific restraint is violated						
	Restraint 1	Restraint 2	Restraint 3	Restraint 4	Restraint 5	Restraint 6	Restraint 7
1	0.717	0.739	0.726	0.729	0.620	0.648	0.656
2	0.863	0.641	0.596	0.576	0.507	0.465	0.483
3	0.855	0.585	0.569	0.572	0.431	0.421	0.418
4	0.077	0.481	0.442	0.456	0.450	0.438	0.417
5	0.319	0.346	0.055	0.137	0.211	0.110	0.023
6	0.227	0.232	0.003	0.005	0.121	0.033	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table S6.9 Total number of accessible complex conformations per number of complying restraints of the PRE5-PUP2 complex using a coarse rotational search (9.72°, 7416 orientations) and grid (2Å).

Number of consistent restraints (N)	Number of accessible complexes consistent with exactly N restraints	Fraction of accessible complexes consistent with exactly N restraints	Number of accessible complexes consistent with at least N restraints	Fraction of accessible complexes consistent with at least N restraints
0	96048647	0.7871	122031044	1.0000
1	9884350	0.0810	25982397	0.2129
2	3940305	0.0323	16098046	0.1319
3	10681417	0.0875	12157741	0.0996
4	1281763	0.0105	1476323	0.0121
5	81860	0.0007	194559	0.0016
6	70057	0.0006	112699	0.0009
7	42641	0.0003	42641	0.0003

Table S6.10 Normalized occurrence of a restraint violation given a number of consistent restraints for the PRE5-PUP2 complex using a coarse rotational search (9.72°, 7416 orientations) and grid (2Å).

Number of consistent restraints (N)	Fraction of complexes consistent with N restraints in which a specific restraint is violated						
	Restraint 1	Restraint 2	Restraint 3	Restraint 4	Restraint 5	Restraint 6	Restraint 7
1	0.713	0.736	0.724	0.726	0.628	0.654	0.662
2	0.861	0.635	0.590	0.570	0.515	0.471	0.489
3	0.853	0.578	0.561	0.565	0.439	0.428	0.426
4	0.076	0.476	0.438	0.452	0.457	0.443	0.422
5	0.320	0.340	0.056	0.136	0.218	0.110	0.022
6	0.227	0.229	0.004	0.005	0.125	0.032	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table S6.11 Profiling disvis for a 20.83° rotational search (648 orientations) using a 1Å grid spacing of RNA polymerase II large subunit and 27kDa polypeptide.

Function	Time (s)	Percentage of total (%)
Determining consistent distance restraint space	7	1
Flattening arrays (method 'flatten' of 'numpy.ndarray' objects)	8	1
Filling arrays (method 'fill' of 'numpy.ndarray' objects)	9	1
Counting violations	11	2
Reduce (method 'reduce' of 'numpy.ufunc' objects)	12	2
Complex conjugate (method 'conj' of 'numpy.ndarray' objects)	19	3
Rotating the scanning chain	24	3
Copying of arrays (method 'copy' of 'numpy.ndarray' objects)	36	5
Binning the number of accessible complexes	37	5
Main loop (multiplications, summations, etc.)	104	15
FFT calculations	445	62
Total time	717	100

6

Chapter 7

Inferring interface residues from the accessible interaction space defined by distance restraints to improve HADDOCKing models

7.1 Introduction

Uncovering the precise atomic structures of protein complexes is a highly sought-after enterprise. Experimental techniques that provide atomic resolution, mainly X-ray crystallography and NMR spectroscopy, have, unfortunately, so far only revealed a fraction of the whole interactome, the set of all interacting proteins [3]. Protein-protein docking aims to predict the structure of a complex from its individual proteins to close this knowledge gap [9]. However, its success rate using solely first-principles – the so-called ab initio docking - is generally low [12]. Integrating additional information (if reliable) during the docking process can increase the confidence in the resulting models, especially when knowledge about the location of the interface is available [14].

Cross-links coupled with mass spectrometry (CXMS) is an upcoming and promising biochemical method that provides inter-residue distance restraints [35, 145]. Multiple chemistries are being developed, making the approach more robust and increasing the information content [156]. Several software packages have already been developed for visualizing cross-links, and calculating their path length [153, 158, 159]. In the previous chapter we introduced DisVis to quantify and visualize the information content of distance restraints. However, interpreting multiple long-range

distance restraints between components from a structural perspective and deducing the interaction surface remains tedious. A simpler interpretation is gained if interface residues can be deduced from the data, as these map directly onto the individual chains and offer a straightforward prediction of the active site. In addition, in the development of protein-protein inhibitors, mainly the protein interfaces are of importance, and less so the precise complex' structure [7]. Such interface information might be useful in complementing the CXMS based distance restraints, since the allowed distance ranges for the latter can be relatively wide (up to 30Å [154]).

Inclusion of CXMS based distance restraints has already been shown to improve the modeling of both proteins and protein complexes with the Rosetta software [47], and to heavily decrease the number of accessible conformations of a complex (see **Chapter 6**). Our data-driven docking software HADDOCK is capable of directly incorporating distance restraints during the docking [42, 43]. Currently, Mass Spec Studio provides an advanced software platform for integrative modeling from MS data, such as hydrogen/deuterium exchange and cross-links, with HADDOCK [160]. However, no thorough benchmark study has been performed to measure the impact and effectiveness of incorporating cross-link based distance restraints in HADDOCK.

Here we introduce a method to infer interface residues when distance restraints are available in addition to models or structures of the components. The method is benchmarked on 90 complexes taken from the Protein-Protein Docking Benchmark 4.0 (PPDB4.0) [118] for 3, 5 and 7 cross-links, respectively, with an upper distance restraint of 30Å, comparable to the information content that is provided by disuccinimidyl suberate (DSS) and bis-sulfosuccinimidyl-suberate (BS3) cross-links [154]. Finally, we show how this can be combined with HADDOCK to complement unambiguous distance restraints by derived interface information, benchmarking it on 24 cases of the PPDB4.0.

7.2 Methods

7.2.1 Inferring interface residues from distance restraints

In the previous Chapter, we have introduced the concept of the accessible interaction space, the set of all possible complexes that are

consistent with a given number of distance restraints. Indeed, the presence of distance restraints between two interacting macromolecular biomolecules can significantly reduce their accessible interaction space. To infer residues that are likely to be at the interface, we assume that these residues are often found to be interacting in the interaction space consistent with the restraints. Important residues may be determined by performing a full-exhaustive 6 dimensional search of the three translational and three rotational degrees of freedom and counting the number of interactions that each solvent accessible residue forms in complexes consistent with a given number of restraints. We define two residues to be interacting when their C_α – C_α distance is smaller than 10Å. We only consider the C_α-atoms of solvent accessible residues of both chains to make the computations more tractable as the number of possible interactions scales with A^2 with A the number of atoms involved. The average number of interactions per complex (AIC) that a residue i forms is given by

$$\overline{N}_i = \frac{\sum_R^{\mathbf{P}} w_R \sum_C^{\mathbf{C}_R} I_C}{\sum_R^{\mathbf{P}} w_R \mathbf{C}_R} \tag{7.1}$$

where the first summation is over all rotations $\mathbf{P}$ indexed by R; w_R is a weight factor to correctly average over rotation space; the second summation is over all complexes $\mathbf{C}_R$ that are formed within a translational scan indexed by C; and I_C is the number of interactions that are formed by residue i in each sampled complex C.

This approach has been implemented in DisVis (see **Chapter 6**, `https://github.com/haddocking/disvis`), which requires for the interaction analysis an extra input file containing the solvent accessible residue numbers for the fixed and scanning chain. As a result, DisVis outputs a file containing the number of interactions that are formed by each residue for complexes consistent with at least N restraints.

7.2.2 Benchmarking interface residue extraction

We benchmarked our approach on 90 complexes taken from the PPDB4.0, of which 58 were classified as Easy, 14 as Medium, and 18 as Difficult. Virtual cross-links were calculated using a local version of the XWalk software [153] on the bound complex. The virtual cross-links

were chosen such that the solvent accessible surface (SAS) distance was shorter or equal than 34Å [47], and the Euclidean distance smaller or equal than 30Å [154] and the cross-linked residues should be present in both the bound and unbound proteins. The cross-links were randomly picked from the list of all virtual cross-links using the SAS-distance dependent probability distribution as was used by Kahraman et al. [47] to mimic experimental cross-link data: 0 – 10Å 9%; 10 – 15Å 18%; 15 – 20Å 34%; 20 – 25Å 22%; and 25 – 34Å 16%.

The solvent accessible residues were determined by running *naccess* [161] on the two unbound proteins. Residues that had a relative solvent accessibility of the main or side chain of 50% or higher were used as surface residues. DisVis runs were performed for each complex using 3, 5, and 7 random restraints, respectively, with a 5.27° rotational sampling, and default values for the voxel spacing (1Å), maximum clashing volume (200Å^3) and minimum interaction volume (300Å^3). The AIC was only calculated from the complexes consistent with all restraints.

Correct interface residues were taken from the experimental structure of the complex using the above definition of interaction, under the restriction that the residue was also present in the unbound proteins. To analyze the predictive capabilities the precision P and recall R were calculated as

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7.2}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7.3}$$

where TP, FP and FN stand for True Positive, False Positive, and False Negative, respectively.

7.2.3 HADDOCKing with virtual cross-links

To determine whether the inclusion of DisVis-determined interface residues aids the docking process of HADDOCK, we benchmarked HADDOCK using 24 complexes of the PPDB4.0, of which 16 were the same as those used by Kahraman et al. [47]. The remaining 8 were randomly picked Easy complexes, as the previous 16 already consisted of 7 Medium

and 9 Difficult cases. HADDOCK was benchmarked with 4 different protocols:

- using the restraints directly as unambiguous distance restraints (unambig) with a minimal and maximal Euclidean length of 0 and 30Å, respectively;
- using the unambiguous restraints in combination with center-of-mass restraints [64];
- using solely DisVis-based ambiguous interaction restraints (AIRs);
- and using a combination of the unambig restraints and DisVis-based AIRs.

Each protocol was performed with 3, 5 and 7 generated virtual cross-link restraints, respectively, as described in the previous section. The DisVis-based AIRs were determined as follows: a DisVis run was performed as described above using the unbound structures of the complex together with the virtual cross-links; active residues were chosen such that their AIC consistent with all restraints had to be larger than 1; passive residues will be chosen with a to be determined AIC cutoff based on the analysis that was described in the previous section . Per HADDOCK run 1000 it0-structures were written to file, using 5 trials per file combined with 180° rotated solutions, resulting in 10,000 sampled solutions ($1000 \times 5 \times 2$), of which the 200 best scoring solutions were subjected to the semi-flexible refinement (it1 and itw) (default settings of the server). The solutions were analyzed by calculating the ligand-RMSD (l-RMSD) against the native complex, by first optimally fitting the receptor chain and afterwards calculating the RMSD of backbone atoms of the ligand chain using ProFitV3.1 [132]. Models with an l-RMSD lower than 10Å were considered acceptable.

7.3 Results and discussion

7.3.1 Inferring interface residues from distance restraints

The analysis of the DisVis benchmark is shown in **Figure 7.1** for all complexes and for each difficulty category, by plotting the precision and

recall against the AIC cutoff of a residue. This shows for example that for all complexes, when using 7 cross-links, for residues that have an AIC ≥ 1.0 the precision is approximately 60%, i.e. 60% of all residues with a AIC ≥ 1.0 are true interface residues; the recall at 1.0 AIC is around 40%, meaning that 40% of all true interface residues are still retained in the set of residues satisfying the AIC cutoff condition.
As expected, the precision increases with increasing AIC, while the recall rate drops steadily. Also, both precision and recall rise with the number of available cross-links, reflecting the higher information content of the distance restraints. On average for all complexes, the precision starts at 30% regardless of the number of available restraints, and rises with increasing cutoff to 50, 70 and 80% for 3, 5, and 7 restraints, respectively. For Easy complexes this increases even to 60, 85 and 90%, while for Medium and Difficult complexes the precision is significantly lower, dropping down to 70% for Difficult complexes in the presence of 7 restraints. We attribute the noisy behavior of the precision, especially for Medium complexes, to the smaller number of complexes sampled compared to the number of Easy complexes (14 versus 58).

The recall percentage averaged over all complexes starts around 90%, a consequence of how surface residues were chosen: because of small conformational changes between the bound and unbound chains, residues that are regarded as solvent accessible in the bound form might not be accessible in the unbound form. This also explains the lower starting recall rate of Medium and Difficult complexes, with the latter being lower than 80%, as these more challenging complexes typically exhibit greater conformational change between their bound and unbound form by definition. Not surprisingly, the recall rate decreases steadily with increasing cutoff, as fewer residues will be satisfying the AIC cutoff condition. In contrast to the precision, the recall rate does increase significantly with the inclusion of more restraints for the more challenging categories (Medium + Difficult).

7.3.2 HADDOCKing with virtual cross-links

The HADDOCK benchmark results for the 8 Easy complexes, and the 16 Medium and Difficult complexes are displayed in **Figure 7.2** and **7.3**, respectively, in terms of the structure with the lowest l-RMSD after the water-refinement stage, irrespective of its rank. For the DisVis-based

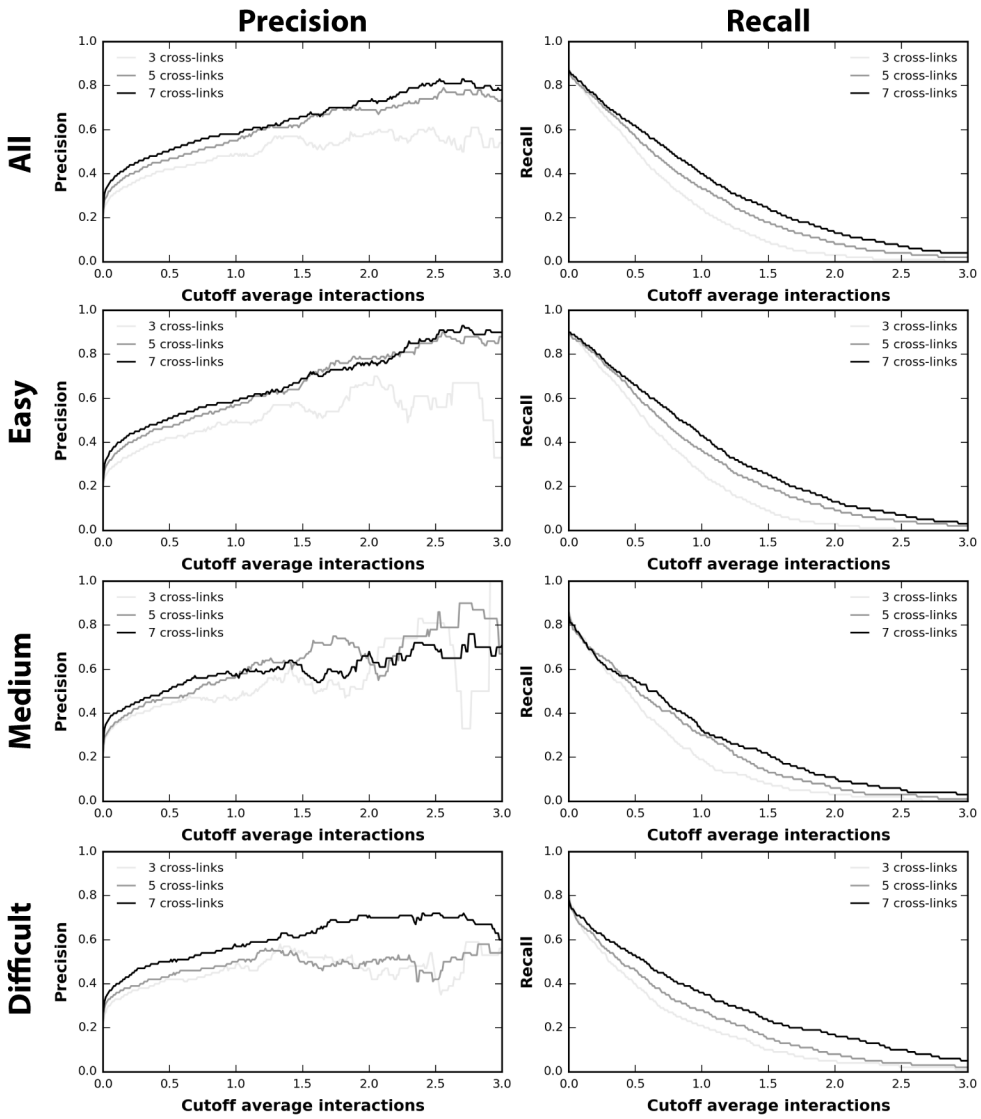

Figure 7.1 Precision and recall rates. The precision and recall rates are plotted against the average-interactions-per-complex (AIC) cutoff. The results are shown averaged over all 90 benchmarked complexes, and each difficulty category (58 Easy, 14 Medium, and 18 Difficult).

AIRs the cutoff AIC for active residues was set to 1, corresponding to a precision of 40 to 60% and a recall of 20 to 40% for 3 and 7 cross-links; the AIC cutoff for passive residues was set to 0.1, as the precision increase is steeper approximately until that point, while still keeping a reasonable recall rate of approximately 80%. For the Easy complexes the unambiguous restraints approach was successful in 50, 37.5 and 75% of the cases when using 3, 5 and 7 cross-links. Adding the center-of-mass restraint this changed to 62.5, 50, and 62.5%. Using the DisVis-based AIRs resulted in a success rate of 25, 62.5, and 75%, and combined with the unambiguous restraints this increased to 50, 75 and 75% success rate. Interestingly, increasing the number of cross-links does not necessarily result in better structures when using only the unambiguous restraints: the success rate with 5 cross-links is markedly lower than using 3. Also, strangely, the unambiguous approach with center-of-mass restraint is the only method for which no acceptable solutions are generated for the 1QA9 complex, even with 7 restraints included. The results of the DisVis-based approaches, however, are improving with increasing numbers of cross-links.

In the 16 Medium and Difficult complexes the unambiguous approach is successful in 18.75, 62.5 and 50% of the cases for 3, 5 and 7 cross-links, respectively; with inclusion of the center-of-mass restraint this becomes 37.5, 56.25, and 75%. The DisVis-based AIRs result in 12.5, 50 and 50% successful cases; including the unambiguous restraints this increases to 12.5, 56.25 and 62.5%. As with the Easy complexes, the success rate of the unambiguous restraints shows no steady improvement with an increased number of cross-links. Furthermore, the combination of DisVis-based AIRs and unambiguous restraints is superior in general to using only DisVis-based AIRs.

The HADDOCK results using 7 cross-links are compared against Rosetta in **Figure 7.3**C. Rosetta was successful in 68.75% of the cases, a slightly higher percentage than HADDOCK using DisVis-based AIRs with unambiguous restraints, but lower than when using unambiguous with center-of-mass restraints. However, in general the results are comparable.

Taking all 24 complexes together we conclude that using only the unambiguous restraints results in a success rate (again defined as generating at least one native-like model in the set of 200 refined models) of 29, 54 and 58% for 3, 5 and 7 cross-links, respectively; using ambiguous restraints with center-of-mass restraints this increases to 46, 54 and 71%.

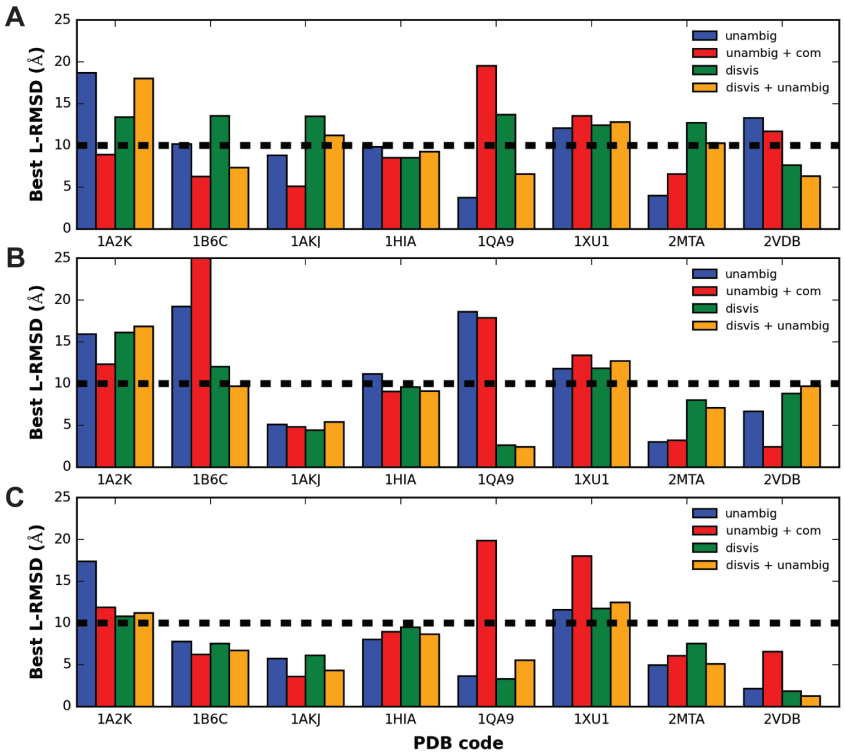

Figure 7.2 Benchmark results on Easy complexes. For each complex the best structure is plotted in terms, of the l-RMSD, for the four procedures tested using (A) 3, (B) 5, and (C) 7 cross-links. The dotted line is the ligand-RMSD cutoff for an acceptable model. Unambig: unambiguous distance restraints; unambig + com: unambiguous restraints combined with center-of-mass restraints; disvis: DisVis-based ambiguous interaction restraints (AIRs); disvis + unambig: unambiguous restraints combined with DisVis-based AIRs.

Using solely DisVis-based AIRs the success rates are respectively 17, 54 and 58%; and DisVis-based AIRs combined with unambiguous restraints 25, 63 and 67%. Based on these results we conclude that if only 3 cross-links are available, the best protocol to use is to combine unambiguous restraints with center-of-mass restraints. If 5 or more cross-links are available it is best to either again combine the unambiguous restraints with the center-of-mass restraints, or combine them with DisVis-based AIRs.

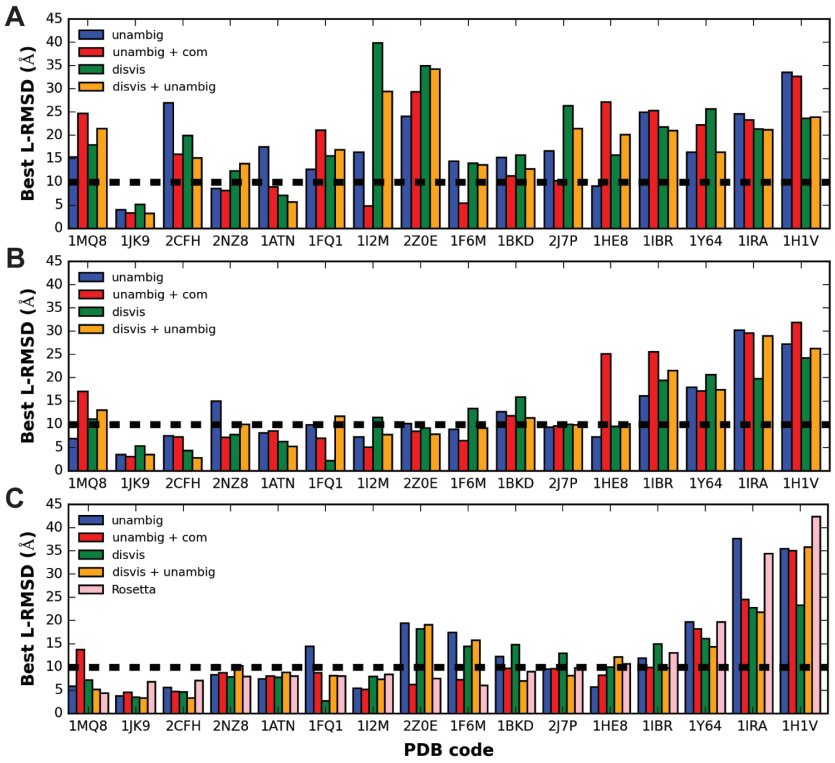

Figure 7.3 Benchmark results on Medium and Difficult complexes. For each complex the best structure is plotted in terms, of the l-RMSD, for the four procedures tested using (A) 3, (B) 5, and (C) 7 cross-links. The dotted line is the ligand-RMSD cutoff for an acceptable model. Unambig: unambiguous distance restraints; unambig + com: unambiguous restraints combined with center-of-mass restraints; disvis: DisVis-based ambiguous interaction restraints (AIRs); disvis + unambig: unambiguous restraints combined with DisVis-based AIRs. In (C) also the best structures of the Rosetta benchmark are shown [47].

7.4 Conclusions

In this Chapter we have introduced a method to infer interface residues by enumerating all interactions a residue forms in the interaction space consistent with all restraints and normalizing it against the number of accessible complexes. It was shown that residues with a higher AIC are more likely to be interface residues with precision reaching almost 90% for rigid complexes. This information can be used to guide future mutagenesis studies and map out the interface, irrespective of the

quality of the generated models of the complex. In addition, we benchmarked several protocols within HADDOCK that incorporated cross-link based distance restraints. Based on the analysis, using the cross-links directly as unambiguous restraints is sub-optimal, and instead should be complemented with either center-of-mass restraints or DisVis-based AIRs. Furthermore, we have shown that HADDOCK and Rosetta are very comparable in their docking performance when including the distance restraints. However, it should be noted that we only investigated the quality of the best model. Further analysis should also address the ranking of generated models, as using various types of restraints might significantly affect the scoring functions' ability to identify near-native models.

7

Chapter 8

Summary and perspectives

8.1 Summary

The previous Chapters in this thesis have introduced and showcased novel approaches for explorative and integrative modeling in the presence of cryo-EM data and distance restraints. In **Chapter 2** I presented the PowerFit software, a Python package for fast cross correlation based rigid body fitting of high-resolution structures in low-resolution densities. PowerFit comes with a new more sensitive scoring function, the core-weighted local cross correlation, in addition to an optimized protocol for fast fitting. In **Chapter 3** I reported results of an extensive benchmark of the PowerFit software using 379 subunits of 5 ribosome density maps. The success rate of unambiguously fitting subunits larger than 100 residues reached approximately 90% up to 12Å resolution, showing that objective fitting methods have matured to usable aids in structural modeling. The limits of rigid body fitting can be leveraged through the use of image pyramids to gain a speedup of a factor of 30 on CPUs and 40 on GPUs, and it allows the identification of possible over-interpreted regions of the density on an objective basis.

Chapter 4 describes the incorporation and benchmarking of cryo-EM data into the data-driven docking program HADDOCK. The approach is flexible and can be fully combined with other available sources of data in HADDOCK, making it a fully integrative modeling approach. It was demonstrated on two ribosome systems, two virus-antibody systems, and a symmetric pentamer. An update of the HADDOCK web server was presented in **Chapter 5**, together with extensive usage statistics of the software all over the world.

Chapter 6 dealt with explorative modeling using distance restraints in general, and cross-link data specifically. I introduced the concept of the accessible interaction space and presented a method to quantify and visualize it. This directly indicates the information content of distance restraints and shows whether all data are self-consistent and, if not, it gives an indication of which restraint is a false-positive. This was implemented in another Python package, DisVis. The approach is general and can easily be incorporated into FFT-based docking programs allowing the use of distance restraints by combining the 'marriage made in heaven' of sampling and scoring [147].

I extended this approach further in **Chapter 7**, presenting a method to infer interface residues from distance restraints using the concept of the average-interactions-per-complex (AIC) statistic. The AIC provides an objective probability for a residue to be at the interface based on the available data. Furthermore, I benchmarked the use of cross-link based distance restraints in HADDOCK using four different approaches. My results show that using solely unambiguous distance restraints is suboptimal; instead they should either be complemented with center-of-mass restraints or DisVis-based ambiguous interactions restraints.

8.2 Challenges of integrative modeling

The field of integrative modeling is still relatively young, with several challenges ahead that the structural biology community will have to face, since integrative approaches are increasingly applied to solve the structure of large macromolecular assemblies. Recently a task force was assigned by the Worldwide PDB (wwPDB) to make recommendations for the field to follow in order to consistently progress and allow a proper assessment of the quality of such integrative models. The results of the First wwPDB Hybrid/Integrative Methods Task Force Workshop were recently published [144], with 5 main recommendations about data-representation, model validation and data-archives. These were that: 1) the experimental and computational protocols in addition to the structural models should be deposited; 2) multiple model representations should be allowed for multi-scale and multi-temporal models; 3) new procedures should be developed to ascertain model uncertainty and accuracy; 4) a federated system of data archives should be created; and

5) publications standards need to be developed for integrative models as is already the case for X-ray and NMR structures.

Thus point 1, 4 and 5 are mainly about the reproducibility of integrative structural models, point 2 is about what data-structures and format standards to use, so far all more practical matters reflecting the current immature status of the field than real inherent scientific challenges. Point 3 highlights a current challenge in this field with respect to the precision and accuracy together with the validation of integrative models. Even though for several experimental techniques, cross-validation (SAXS [162]) and confidence interval (cryo-EM [89]) measures have been developed, they have been infrequently used, except in X-ray crystallography where this has been a standard since years (the concept of the free R-factor [163]), and thus far not been combined. For other methods such as cross-links coupled with mass-spectrometry (CXMS) the statistical propensities of derived distance restraints have only been sparsely studied for small benchmark and sample sizes [47, 156].

Gaining deeper insight into the uncertainty of integrative models and current validation approaches, requires new high-quality and elaborate benchmarks on systems for which high-resolution structures of both the bound and unbound states are available, of which the protein-protein docking benchmark is a prime example [164] (although not really representative of the complexity of systems typically studies by integrative modeling approaches), together with additional experimental data. Especially for upcoming promising techniques as SAXS and CXMS, experimental data on multiple well-investigated systems are missing even for binary protein interactions. Although there are databases for CXMS [47, 165], they are relatively limited in size, e.g. the XLdb reports 62 intra-chain cross-links of which 34 are coming from a single RNA polymerase II system [47, 145]. The small sample size and questionable reproducibility of the results are major limitations in the development of robust validation and uncertainty assessment tools. Thus, for the integrative structural biology field to properly move forward a *quid pro quo* mentality needs to be established between experimental and computational scientists: additional experiments should be performed for the purpose of further understanding the scope and limitations that the data are providing, so that, in turn, improved computational models can be delivered to answer important biological questions.

8.3 Future guidelines and additional fields of research

8.3.1 Explorative modeling

To adequately model the uncertainty of integrative models more emphasis should be put on the data themselves by investigating the amount of information the data carry by searching and quantifying the whole interaction space. I presented in this thesis a methodology in **Chapter 6** to assess the information content of distance restraints, information that can be obtained from a variety of techniques. Note, however, that the approach presented is limited to binary complexes and fully characterizing the interaction space of multi-component systems remains an open challenge. The approach for appreciating the information content of distance restraints can be further extended by using a statistical distance preference function, i.e. a knowledge based potential, inferred from experimental data, to better investigate the probability distribution of the accessible interaction space. Similar approaches can be developed for SAXS (though computationally more expensive as the scattering curve needs to be calculated millions to billions of times, a more CPU-demanding process than a simple distance calculation), and other biochemical and biophysical based potentials, such as surface overlap/van der Waals interactions. Thus, instead of heuristically optimizing the number of acceptable models within the top X best scoring structures using a linear combination of (pseudo-)energies, as is common in the docking field [166], the energy distributions can be analyzed to give further indication of the reliability of each measure and from there to define confidence intervals in models. Established probability distributions can afterwards be used as Bayesian priors in an effort to move to Bayesian statistical models.

Furthermore, current integrative modeling is often used to generate only a handful or even, preferably, a single representative model of the data, even though the original outset of the approach is to generate all data-consistent models. This is unfortunate, since it hides many nuances and complexities of the biological systems. For the integrative approach to live up to its potential, requires a different mindset of the structural biologist in general: a structural model should not be simply regarded as a single entity, but rather as a whole set of conformations, as is already the case in NMR structure ensembles. This insight is now also gaining ground in X-ray crystallography, where methods are being developed

that represent the electron density as a set of conformers [167, 168] and ensembles [169]. These representations are only the tip of the iceberg within this mindset, as the ensemble space will be significantly bigger in the presence of sparse data, such as CXMS data. Model representations should thus become more diffuse with larger accessible interaction spaces, to accurately present the ensembles consistent with the data. Again, explorative modeling techniques can help here by quantifying the information content to provide insight into the magnitude of the ensemble space, while concurrently easing the transition from a single-structure mindset to an elaborate multi-ensemble paradigm.

8.3.2 Formal structural biology

Further investigating the accuracy of individual experimental methods require scientists that are trained in both computational and experimental techniques, the *hybrid scientist*. This allows the scientist to perform experiments to further guide and validate the computational modeling, ultimately resulting in a *formal structural biology*, where instead of only advancing biological insight, the emphasis is also put on investigating the accuracy and precision of both models and experimental methods *an sich* and the interpretation of the generated results. In the semi-long run, this approach will become a fertile and stable foundation to build upon for in-depth structural research in challenging and interesting biological systems and networks. This will ultimately result in a more formal approach to structure determination from multiple data sources.

8.4 Conclusion

The (integrative) structural biology field is a fast moving and exciting field of research, with many experimental and computational advances. The most recent dramatic example is of course the spectacular improvement of the cryo-EM field due to direct electron detectors. Even though atomic resolution can now be achieved for stable complexes, the bulk of the resulting densities need additional data from diverse sources for a structural interpretation, requiring high-end integrative methods. However, there are still significant challenges to overcome for integrative

modeling to become a standard tool in the toolbox of structural biologists. These are mainly dealing with the reproducibility of the results and the uncertainty of the models. By showcasing integrative modeling approaches and introducing new methods for quantifying the information content of experimental data this thesis has laid out some new building blocks for the field to build upon and move forward.

References

1 Nietzsche, F. (1891). Also sprach Zarathustra: Ein Buch für Alle und Keinen. Ernst Schmeitzner.

2 Braun, P. and Gingras, A.-C. (2012). History of protein-protein interactions: from egg-white to complex networks. *Proteomics* **12**, 1478–1498.

3 Stein, A., Mosca, R. and Aloy, P. (2011). Three-dimensional modeling of protein interactions and complexes is going 'omics. *Curr Opin Struct Biol* **21**, 200–208.

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

5 Campbell, I. D. (2002). Timeline: the march of structural biology. *Nat Rev Mol Cell Biol* **3**, 377–381.

6 Bienstock, R. J. (2012). Computational drug design targeting protein-protein interactions. *Curr Pharm Des* **18**, 1240–1254.

7 Sable, R. and Jois, S. (2015). Surfing the Protein-Protein Interaction Surface Using Docking Methods: Application to the Design of PPI Inhibitors. *Molecules* **20**, 11569–11603.

8 Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R. and Melo, F. et al. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29**, 291–325.

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

10 Kundrotas, P. J., Zhu, Z., Janin, J. and Vakser, I. A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proc Natl Acad Sci U S A* **109**, 9438–9441.

11 Vakser, I. A. (2013). Low-resolution structural modeling of protein interactome. *Curr Opin Struct Biol* **23**, 198–205.

12 Huang, S.-Y. (2015). Exploring the potential of global protein-protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug Discov Today* **20**, 969–977.

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

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

15 Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W. and Kipper, J. et al. (2007a). Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694.

16 Ward, A. B., Sali, A. and Wilson, I. A. (2013). Biochemistry. Integrative structural biology. *Science* **339**, 913–915.

17 Schneidman-Duhovny, D., Pellarin, R. and Sali, A. (2014). Uncertainty in integrative structural modeling. *Curr Opin Struct Biol* **28**, 96–104.

18 Case, D. A. (2013). Chemical shifts in biomolecules. *Curr Opin Struct Biol* **23**, 172–176.

19 Chen, K. and Tjandra, N. (2012). The use of residual dipolar coupling in studying proteins by NMR. *Top Curr Chem* **326**, 47–67.

20 van Ingen, H. and Bonvin, A. M. J. J. (2014). Information-driven modeling of large macromolecular assemblies using NMR data. *J Magn Reson* **241**, 103–114.

21 Putnam, C. D., Hammel, M., Hura, G. L. and Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys* **40**, 191–285.

22 Schneidman-Duhovny, D., Kim, S. J. and Sali, A. (2012). Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct Biol* **12**, 17.

23 Blanchet, C. E. and Svergun, D. I. (2013). Small-angle X-ray scattering on biological macromolecules and nanocomposites in solution. *Annu Rev Phys Chem* **64**, 37–54.

24 Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G. and Kohlbacher, O. et al. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430.

25 Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H. and Earl, L. A. et al. (2013). Cryo-electron microscopy – a primer for the non-microscopist. *FEBS J* **280**, 28–45.

26 Glaeser, R. M. and Taylor, K. A. (1978). Radiation damage relative to transmission electron microscopy of biological specimens at low temperature: a review. *J Microsc* **112**, 127–138.

27 Smith, M. T. J. and Rubinstein, J. L. (2014). Structural biology. Beyond blob-ology. *Science* **345**, 617–619.

28 Bai, X.-c., McMullan, G. and Scheres, S. H. W. (2015). How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* **40**, 49–57.

29 Nogales, E. and Scheres, S. H. W. (2015). Cryo-EM: A unique tool for the visualization of macromolecular complexity. *Mol Cell* **58**, 677–689.

30 Gonen, T., Cheng, Y., Sliz, P., Hiroaki, Y. and Fujiyoshi, Y. et al. (2005). Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. *Nature* **438**, 633–638.

31 Schur, F. K. M., Hagen, W. J. H., de Marco, A. and Briggs, J. A. G. (2013). Determination of protein structure at 8.5Å resolution using cryo-electron tomography and sub-tomogram averaging. *J Struct Biol* **184**, 394–400.

32 Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D. and Wu, X. et al. (2015). Electron microscopy. 2.2 Å resolution cryo-EM structure of -galactosidase in complex with a cell-permeant inhibitor. *Science* **348**, 1147–1151.

33 Baker, M. L., Baker, M. R., Hryc, C. F. and Dimaio, F. (2010). Analyses of subnanometer resolution cryo-EM density maps. *Methods Enzymol* **483**, 1–29.

34 Tran, B. Q., Goodlett, D. R. and Goo, Y. A. (2015). Advances in protein complex analysis by chemical cross-linking coupled with mass spectrometry (CXMS) and bioinformatics. *Biochim Biophys Acta*.

35 Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F. and Rinner, O. et al. (2010). Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol Cell Proteomics* **9**, 1634–1649.

36 Merkley, E. D., Cort, J. R. and Adkins, J. N. (2013). Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze. *J Struct Funct Genomics* **14**, 77–90.

37 Tosi, A., Haas, C., Herzog, F., Gilmozzi, A. and Berninghausen, O. et al. (2013). Structure and subunit topology of the INO80 chromatin remodeler and its nucleosome complex. *Cell* **154**, 1207–1219.

38 Ciferri, C., Lander, G. C., Maiolica, A., Herzog, F. and Aebersold, R. et al. (2012). Molecular architecture of human polycomb repressive complex 2. *Elife* **1**, e00005.

39 Erzberger, J. P., Stengel, F., Pellarin, R., Zhang, S. and Schaefer, T. et al. (2014). Molecular architecture of the 40SeIF1eIF3 translation initiation complex. *Cell* **158**, 1123–1135.

40 Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F. and Thompson, J. et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545–574.

41 Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J. and Tjioe, E. et al. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* **10**, e1001244.

42 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.

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

44 Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. and Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66–93.

45 Simons, K. T., Bonneau, R., Ruczinski, I. and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **Suppl 3**, 171–176.

46 DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A. and Oberdorfer, G. et al. (2011). Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **473**, 540–543.

47 Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G. and Aebersold, R. et al. (2013). Cross-link guided molecular modeling with ROSETTA. *PLoS One* **8**, e73411.

48 Demers, J.-P., Habenstein, B., Loquet, A., Kumar Vasa, S. and Giller, K. et al. (2014). High-resolution structure of the Shigella type-III secretion needle by solid-state NMR and cryo-electron microscopy. *Nat Commun* **5**, 4976.

49 DiMaio, F., Song, Y., Li, X., Brunner, M. J. and Xu, C. et al. (2015). Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat Methods* **12**, 361–365.

50 Adams, P. D., Baker, D., Brunger, A. T., Das, R. and DiMaio, F. et al. (2013). Advances, interactions, and future developments in the CNS, Phenix, and Rosetta structural biology software systems. *Annu Rev Biophys* **42**, 265–287.

51 Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W. and Kipper, J. et al. (2007b). The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701.

52 Webb, B., Lasker, K., Schneidman-Duhovny, D., Tjioe, E. and Phillips, J. et al. (2011). Modeling of proteins and their assemblies with the integrative modeling platform. *Methods Mol Biol* **781**, 377–397.

53 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.

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

55 Schneidman-Duhovny, D., Hammel, M. and Sali, A. (2010). FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res* **38**, W540–W544.

56 Yang, Z., Lasker, K., Schneidman-Duhovny, D., Webb, B. and Huang, C. C. et al. (2012). UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J Struct Biol* **179**, 269–278.

57 van Dijk, A. D. J., Fushman, D. and Bonvin, A. M. J. J. (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.

58 van Dijk, A. D. J., Kaptein, R., Boelens, R. and Bonvin, A. M. J. J. (2006a). Combining NMR relaxation with chemical shift perturbation data to drive protein-protein docking. *J Biomol NMR* **34**, 237–244.

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

60 van Dijk, A. D. J. and Bonvin, A. M. J. J. (2006). Solvated docking: introducing water into the modelling of biomolecular complexes. *Bioinformatics* **22**, 2340–2347.

61 Kastritis, P. L., van Dijk, A. D. J. and Bonvin, A. M. J. J. (2012). Explicit treatment of water molecules in data-driven protein-protein docking: the solvated HADDOCKing approach. *Methods Mol Biol* **819**, 355–374.

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

63 Schmitz, C. and Bonvin, A. M. J. J. (2011). Protein-protein HADDocking using exclusively pseudocontact shifts. *J Biomol NMR* **50**, 263–266.

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

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

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

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

68 Villa, E. and Lasker, K. (2014). Finding the right fit: chiseling structures out of cryo-electron microscopy maps. *Curr Opin Struct Biol* **25**, 118–125.

69 Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S. and Greenblatt, D. M. et al. (2004). UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612.

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

71 Volkmann, N. and Hanein, D. (1999). Quantitative fitting of atomic models into observed densities derived by electron microscopy. *J Struct Biol* **125**, 176–184.

72 Roseman, A. M. (2000). Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Crystallogr D Biol Crystallogr* **56**, 1332–1340.

73 Chacón, P. and Wriggers, W. (2002). Multi-resolution contour-based fitting of macromolecular structures. *J Mol Biol* **317**, 375–384.

74 Kovacs, J. A., Chacón, P., Cong, Y., Metwally, E. and Wriggers, W. (2003). Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Crystallogr D Biol Crystallogr* **59**, 1371–1376.

75 Wu, X., Milne, J. L. S., Borgnia, M. J., Rostapshov, A. V. and Subramaniam, S. et al. (2003). A core-weighted fitting method for docking atomic structures into low-resolution maps: application to cryo-electron microscopy. *J Struct Biol* **141**, 63–76.

76 Garzón, J. I., Kovacs, J., Abagyan, R. and Chacón, P. (2007). ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics* **23**, 427–433.

77 Hrabe, T., Chen, Y., Pfeffer, S., Cuellar, L. K. and Mangold, A.-V. et al. (2012). PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *J Struct Biol* **178**, 177–188.

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

79 Roseman, A. M. (2003). Particle finding in electron micrographs using a fast local correlation algorithm. *Ultramicroscopy* **94**, 225–236.

80 Karney, C. F. F. (2007). Quaternions in molecular modeling. *J Mol Graph Model* **25**, 595–604.

81 Anger, A. M., Armache, J.-P., Berninghausen, O., Habeck, M. and Subklewe, M. et al. (2013). Structures of the human and Drosophila 80S ribosome. *Nature* **497**, 80–85.

82 Ranson, N. A., Farr, G. W., Roseman, A. M., Gowen, B. and Fenton, W. A. et al. (2001). ATP-bound states of GroEL captured by cryo-electron microscopy. *Cell* **107**, 869–879.

83 Volkmann, N. (2002). A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. *J Struct Biol* **138**, 123–129.

84 Pintilie, G. and Chiu, W. (2012). Comparison of Segger and other methods for segmentation and rigid-body docking of molecular components in cryo-EM density maps. *Biopolymers* **97**, 742–760.

85 Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. and Gossard, D. C. (2010). Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J Struct Biol* **170**, 427–438.

86 Chen, D.-H., Madan, D., Weaver, J., Lin, Z. and Schröder, G. F. et al. (2013). Visualizing GroEL/ES in the act of encapsulating a folding protein. *Cell* **153**, 1354–1365.

87 Guo, Q., Yuan, Y., Xu, Y., Feng, B. and Liu, L. et al. (2011). Structural basis for the function of a small GTPase RsgA on the 30S ribosomal subunit maturation revealed by cryoelectron microscopy. *Proc Natl Acad Sci U S A* **108**, 13100–13105.

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

89 Volkmann, N. (2009). Confidence intervals for fitting of atomic models into low-resolution densities. *Acta Crystallogr D Biol Crystallogr* **65**, 679–689.

90 Derevyanko, G. and Grudinin, S. (2014). HermiteFit: fast-fitting atomic structures into a low-resolution density map using three-dimensional orthogonal Hermite functions. *Acta Crystallogr D Biol Crystallogr* **70**, 2069–2084.

91 Cyganek, B. and Siebert, J. (2009). An introduction to 3D computer vision techniques and algorithms. John Wiley & Sons, Ltd., Chicester.

92 Lawson, C. L., Baker, M. L., Best, C., Bi, C. and Dougherty, M. et al. (2011). EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res* **39**, D456–D464.

93 Gutmanas, A., Alhroub, Y., Battle, G. M., Berrisford, J. M. and Bochet, E. et al. (2014). PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* **42**, D285–D291.

94 Armache, J.-P., Jarasch, A., Anger, A. M., Villa, E. and Becker, T. et al. (2010). Localization of eukaryote-specific ribosomal proteins in a 5.5-Å cryo-EM map of the 80S eukaryotic ribosome. *Proc Natl Acad Sci U S A* **107**, 19754–19759.

95 Budkevich, T. V., Giesebrecht, J., Behrmann, E., Loerke, J. and Ramrath, D. J. F. et al. (2014). Regulation of the mammalian elongation cycle by subunit rolling: a eukaryotic-specific ribosome rearrangement. *Cell* **158**, 121–131.

96 Aylett, C. H. S., Boehringer, D., Erzberger, J. P., Schaefer, T. and Ban, N. (2015). Structure of a yeast 40S-eIF1-eIF1A-eIF3-eIF3j initiation complex. *Nat Struct Mol Biol* **22**, 269–271.

97 Svidritskiy, E., Brilot, A. F., Koh, C. S., Grigorieff, N. and Korostelev, A. A. (2014). Structures of yeast 80S ribosome-tRNA complexes in the rotated and nonrotated conformations. *Structure* **22**, 1210–1218.

98 Weisstein, E. (2015). MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/Convolution.html.

99 von Loeffelholz, O., Jiang, Q., Ariosa, A., Karuppasamy, M. and Huard, K. et al. (2015). Ribosome-SRP-FtsY cotranslational targeting complex in the closed state. *Proc Natl Acad Sci U S A* **112**, 3943–3948.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

114 Schneidman-Duhovny, D., Rossi, A., Avila-Sakar, A., Kim, S. J. and Velázquez-Muriel, J. et al. (2012). A method for integrative structure determination of protein-protein complexes. *Bioinformatics* **28**, 3282–3289.

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

116 de Vries, S. J., Melquiond, A. S. J., Kastritis, P. L., Karaca, E. and Bordogna, A. et al. (2010). Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. *Proteins* **78**, 3242–3249.

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

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

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

120 Wang, Z., Li, L., Pennington, J. G., Sheng, J. and Yap, M. L. et al. (2013). Obstruction of dengue virus maturation by Fab fragments of the 2H2 antibody. *J Virol* **87**, 8909–8915.

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

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

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

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

125 Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D. A. and Adams, C. M. et al. (2013). Architecture of an RNA polymerase II transcription pre-initiation complex. *Science* **342**, 1238724.

126 Janin, J., Henrick, K., Moult, J., Eyck, L. T. and Sternberg, M. J. E. et al. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **52**, 2–9.

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

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

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

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

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

132 Martin, A. (2009). ProFitV3.1 http://www.bioinf.org.uk/software/profit/.

133 Ashkenazy, H., Erez, E., Martz, E., Pupko, T. and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529–W533.

134 Rodrigues, J. P. G. L. M., Melquiond, A. S. J., Karaca, E., Trellet, M. and van Dijk, M. et al. (2013). Defining the limits of homology modeling in information-driven protein docking. *Proteins* **81**, 2119–2128.

135 de Vries, S. J. and Bonvin, A. M. J. J. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One* **6**, e17695.

136 Janin, J. (2005). Assessing predictions of protein-protein interaction: the CAPRI experiment. *Protein Sci* **14**, 278–283.

137 Lensink, M. F. and Wodak, S. J. (2013). Docking, scoring, and affinity prediction in CAPRI. *Proteins* **81**, 2082–2095.

138 Wassenaar, T., van Dijk, M., Loureiro-Ferreira, N., van der Schot, G. and de Vries, S. et al. (2012). WeNMR: Structural Biology on the Grid *J Grid Comput* **10**, 743-767.

139 Chen, V. B., Arendall 3rd, W. B., Headd, J. J., Keedy, D. A. and Immormino, R. M. et al. (2010a). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12–21.

140 Kastritis, P. L., Visscher, K. M., van Dijk, A. D. J. and Bonvin, A. M. J. J. (2013). Solvated protein-protein docking using Kyte-Doolittle-based water preferences. *Proteins* **81**, 510–518.

141 van Dijk, M., Visscher, K. M., Kastritis, P. L. and Bonvin, A. M. J. J. (2013). Solvated protein-DNA docking using HADDOCK. *J Biomol NMR* **56**, 51–63.

142 Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. and Karaca, E. et al. (2012). Clustering biomolecular complexes by residue contacts similarity. *Proteins* **80**, 1810–1817.

143 Escobar-Cabrera, E., Okon, M., Lau, D. K. W., Dart, C. F. and Bonvin, A. M. J. J. et al. (2011). Characterizing the N- and C-terminal Small ubiquitin-like modifier (SUMO)-interacting motifs of the scaffold protein DAXX. *J Biol Chem* **286**, 19816–19829.

144 Sali, A., Berman, H. M., Schwede, T., Trewhella, J. and Kleywegt, G. et al. (2015). Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–1167.

145 Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J Struct Biol* **173**, 530–540.

146 Kalinin, S., Peulen, T., Sindbert, S., Rothwell, P. J. and Berger, S. et al. (2012). A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nat Methods* **9**, 1218–1225.

147 Vajda, S., Hall, D. R. and Kozakov, D. (2013). Sampling and scoring: a marriage made in heaven. *Proteins* **81**, 1874–1884.

148 Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A. and Aflalo, C. et al. (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.

149 Walt, S. v. d., Colbert, S. C. and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation *Comput Sci Eng* **13**, 22-30.

150 Behnel, S., Bradshaw, R., Citro, C., Dalcin, L. and Seljebotn, D. S. et al. (2011). Cython: The Best of Both Worlds *Comput Sci Eng* **13**, 31-39.

151 Stone, J. E., Gohara, D. and Shi, G. (2010). OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems *IEEE Des. Test* **12**, 66–73.

152 Klöckner, A., Pinto, N., Lee, Y., Catanzaro, B. and Ivanov, P. et al. (2012). PyCUDA and PyOpenCL: A Scripting-Based Approach to GPU Run-Time Code Generation *Parallel Comput* **38**, 157–174.

153 Kahraman, A., Malmström, L. and Aebersold, R. (2011). Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics* **27**, 2163–2164.

154 Merkley, E. D., Rysavy, S., Kahraman, A., Hafen, R. P. and Daggett, V. et al. (2014). Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances. *Protein Sci* **23**, 747–759.

155 Kiefer, F., Arnold, K., Künzli, M., Bordoli, L. and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* **37**, D387–D392.

156 Leitner, A., Joachimiak, L. A., Unverdorben, P., Walzthoeni, T. and Frydman, J. et al. (2014). Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. *Proc Natl Acad Sci U S A* **111**, 9455–9460.

157 Chen, Z. A., Jawhari, A., Fischer, L., Buchen, C. and Tahir, S. et al. (2010b). Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J* **29**, 717–726.

158 Holding, A. N., Lamers, M. H., Stephens, E. and Skehel, J. M. (2013). Hekate: software suite for the mass spectrometric analysis and three-dimensional visualization of cross-linked protein samples. *J Proteome Res* **12**, 5923–5933.

159 Kosinski, J., von Appen, A., Ori, A., Karius, K. and Müller, C. W. et al. (2015). Xlink Analyzer: software for analysis and visualization of cross-linking data in the context of three-dimensional structures. *J Struct Biol* **189**, 177–183.

160 Rey, M., Sarpe, V., Burns, K. M., Buse, J. and Baker, C. A. H. et al. (2014). Mass spec studio for integrative structural biology. *Structure* **22**, 1538–1548.

161 Hubbard, S. and Thornton, J. (1992). NACCESS http://www.bioinf.manchester.ac.uk/naccess/.

162 Rambo, R. P. and Tainer, J. A. (2013). Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* **496**, 477–481.

163 Bruenger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475.

164 Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G. and Kastritis, P. L. et al. (2015). Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* **427**, 3031–3041.

165 Zheng, C., Weisbrod, C. R., Chavez, J. D., Eng, J. K. and Sharma, V. et al. (2013). XLink-DB: database and software tools for storing and visualizing protein interaction topology data. *J Proteome Res* **12**, 1989–1995.

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

167 Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P. T. and Holton, J. M. et al. (2011). Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc Natl Acad Sci U S A* **108**, 16247–16252.

168 van den Bedem, H., Bhabha, G., Yang, K., Wright, P. E. and Fraser, J. S. (2013). Automated identification of functional dynamic contact networks from X-ray crystallography. *Nat Methods* **10**, 896–902.

169 Burnley, B. T., Afonine, P. V., Adams, P. D. and Gros, P. (2012). Modelling dynamics in protein crystal structures by ensemble refinement. *Elife* **1**, e00311.

Summary

The remarkable diversity and complexity of life in all its facets is a source of constant wonder and amazement all through the history of humankind. Even though every unique individual has his own experience and interpretation on how to approach life, the scientific inquiry of modern man has resulted in a paradigm that life is organized on the molecular level, where the typical dimension is that of the ångstrom (10^{-10}m). This insight has lead to an intense interest in the molecules of life. With the discovery of the double-helix structure of DNA, the biomolecule that holds the genetic code, as a prime example, it is postulated that function follows structure, i.e. knowledge of the precise three-dimensional structure of large biomolecules gives an indication of their function. Maybe more important, having precise knowledge of the biomolecular structure holds the promise of rational drug design by developing biologically active molecules that specifically interact with particular patches at the interface of a complex or in the active site of an enzyme.

X-ray crystallography and NMR spectroscopy are the classical experimental methods that are capable of elucidating the atomic arrangement of large biomolecules, such as proteins. The importance of proteins in the cell cannot be understated: they are the main actors in almost all cellular processes, ranging from muscle contraction to the building of new proteins through the ribosome. Currently, more than 100,000 structures have been solved and deposited in the Protein DataBank (**https://www.pdbe.org**). However, the vast majority of the structures are single proteins, while proteins typically perform their function through interacting with other biomolecules, resulting in large biomolecular complexes. Since the difficulty of solving a structure depends on several parameters, such as the complex' size, binding strength and environment, and the number of biomolecular complexes is estimated to be two orders of magnitude bigger than the number of individual proteins, complementary computational methods are required to close the structure knowledge gap.

Integrative modeling is a particular approach to computationally predict or model the structure of a biomolecular complex by combining all experimental knowledge that is available for the system. It is assumed that this ultimately results in a more accurate and precise model, than

using a single kind of data. The challenges within this approach are, predictably, sampling and scoring: many different possible models need to be generated (sampling) and evaluated to identify the correct ones (scoring). In this thesis, I mainly focus on incorporating cryo-electron microscopy data (cryo-EM) and chemical cross-links coupled with mass spectrometry (CXMS) in the modeling process. Cryo-EM is a fast developing method that typically provides low-resolution density information of large macromolecular complexes, though, with current advances, near-atomic resolution can be achieved; in contrast, CXMS provides long-range distance restraints between amino acids, each thus providing orthogonal information.

In **Chapter 1** I layout a more technical introduction to integrative modeling and the experimental data used. I introduce major software used today that offer a wide array of integrative methods for macromolecular modeling, such as our in-house data-driven docking software HADDOCK. In addition, I propose a new concept, that of explorative modeling, where the emphasis is put on quantifying and preferably visualizing the information content of the data available, rather than outputting specific data-consistent models as is the case in integrative modeling.

Chapter 2 describes PowerFit, a high-performance software package and program for automatic rigid-body fitting of high-resolution biomolecular structures into low-resolution cryo-EM density maps. PowerFit performs a systematic 6 dimensional search of the 3 translational and 3 rotational degrees of freedom to find local cross-correlation optima to objectively place structures in cryo-EM densities. In addition, I introduce a new and sensitive core-weighted local cross-correlation that further extends the applicable resolution range for successful fitting. PowerFit is a first step in this thesis into combining high-resolution structural data with lower-resolution cryo-EM data.

Next, in **Chapter 3** I quantify the resolution requirements for successfully rigid-body fitting a biomolecular structure into a cryo-EM map as a function of the biomolecule's size. I furthermore unambiguously show that the core-weighted Laplacian-enhanced local cross-correlation function is the best performing score overall. Finally, since the resolution limits for successfully fitting a subunit are often remarkably lower than the resolution of current cryo-EM data, these limits can leveraged by using the concept of multi-scale image pyramids, to significantly accelerate the fitting performance and reduce computational time and resources.

Chapter 4 discusses the incorporation of cryo-EM data in HADDOCK, resulting in a truly integrative modeling approach, allowing their combination with all other data sources already supported in HADDOCK. A central concept in the approach is the use of centroids, points that represent the approximate center of mass of each subunit that is docked. HADDOCK also allows the use here of ambiguous restraints if the location of each chain cannot be differentiated in the density. The use of low-resolution cryo-EM data notably increases both the number and quality of acceptable solutions generated through our approach. The use of this powerful integrative method is demonstrated on two ribosome and two virus systems, where additional details of the interface are revealed, providing new insights to guide future experiments.

The HADDOCK2.2 web server is described in **Chapter 5**. The web server offers structural biologists a user-friendly interface to the upgraded HADDOCK2.2 software. Notable features are the introduction of mixed molecule types, e.g. protein-DNA complexes, and additional NMR-based restraints, such as residual dipolar couplings and pseudocontact shifts, opening up new venues for macromolecular docking. The web server can be used free-of-charge for academic purposes at `http://haddock.science.uu.nl/services/HADDOCK2.2` after simple registration.

Starting from **Chapter 6**, I shift to the use of distance restraints in general and CXMS data in particular. I introduce another software package and program, named DisVis for Distance Visualization. DisVis quantifies and visualizes the information content of distance restraints through the concept of the accessible interaction space, the countable set of all data-consistent solutions. It furthermore allows the identification of possible false-positive restraints, and shows whether all restraints are consistent. DisVis represents a first effort into the newly introduced concept of explorative modeling.

The approach is further extended in **Chapter 7**, where explorative modeling is used to infer interface residues in a model-free approach. It is shown that interface residues can be predicted up to 90% accurate for rigid binders in the presence of 3 to 7 long-range distance restraints. The resulting possible interface patches can subsequently be used in HADDOCK as active and passive residues, to enhance the robustness of the CXMS data by combining it with the standard unambiguous distance restraints, which often are not very accurate in terms of distance ranges in the case of CXMS data.

In the final **Chapter 8** I summarize my findings, present a personal perspective on the field of integrative modeling and propose additional fields for future research. I plea for the education of the *hybrid scientist*, an expert at the interface of experimental and computational sciences, to push forward integrative computational modeling. I also argue for a reinterpretation of structural models, as they should not be regarded as single well-defined structures, but instead as ensembles, as has been proposed for NMR. Especially in integrative modeling where models are typically generated using sparse data, the ensemble space is significantly more diffuse. Explorative modeling and the concept of the countable accessible interaction space can help in easing the transition to this ensemble interpretation, while at the same time quantifying the information content of the data.

Overall, this thesis provides new approaches and computational tools to help structural biologists in interpreting data coming from increasingly diverse sources, and introduces a new vantage point to approach modeling through explorative modeling.

Samenvatting

De opmerkelijke diversiteit en complexiteit van het leven in al zijn vormen is een bron van constante verwondering en verbazing gedurende de gehele menselijke geschiedenis. Ook al heeft ieder uniek individu zijn eigen ervaringen en kijk op hoe het leven te benaderen, wetenschappelijk onderzoek van de moderne mens heeft geresulteerd in een paradigma dat leven is georganiseerd op het moleculare niveau, daar waar de typische afstand de ångstrom (10^{-10}m) is. Dit inzicht heeft geleid naar een intense interesse in de moleculen van het leven. Met de ontdekking van de dubbele-helixstructuur van DNA, het biomolecuul dat de genetische code vasthoudt, als een voornaam voorbeeld, wordt er gepostuleerd dat vanuit structuur de functie volgt, dat wil zeggen dat kennis van de precieze drie dimensionale structuur van grote biomoleculen een indicatie geeft van hun functie. Wat misschien nog belangrijker is, de precieze kennis van de biomoleculaire structuur draagt de belofte voor het rationeel ontwerpen van medicijnen door biologisch actieve moleculen te ontwikkelen die specifieke interacties aangaan met bepaalde delen van het oppervlak van een biomolecuul of in de actieve locatie van een enzym.

Röntgendiffractie en kernspinresonantie (NMR) spectroscopie zijn de klassieke experimentele methodes die de atomaire rangschikking kunnen onthullen van grote biomoleculen, zoals eiwitten. De invloed van eiwitten in de cel kan niet overgewaardeerd worden: zij zijn de voornaamste spelers in bijna ieder cellulair proces, variërend van spiercontractie tot het bouwen van andere eiwitten middels het ribosome. Op dit moment zijn meer dan 100,000 structuren opgelost en geplaatst in de EiwitDataBank (**https://www.pdbe.org**). Echter, de grote meerderheid van deze structuren zijn individuele eiwitten, terwijl eiwitten meestal hun functie uitvoeren door het aangaan van interacties met andere biomoleculen, wat resulteert in grote biomoleculaire complexen. Omdat de ingewikkeldheid van het oplossen van een structuur afhangt van meerdere variabelen, zoals de grote van het complex, de bindingssterkte en mogelijke membraanomgeving, en het aantal biomoleculare complexen wordt geschat op ongeveer 100 keer het aantal van individuele eiwitten, zijn complementaire computationele methoden vereist om het gat in de structuurkennis te verkleinen.

Integratief modelleren is een benadering om de structuur van een biomoleculair complex te verspellen/modelleren door het combineren van alle experimentele kennis die beschikbaar is voor het systeem. Er wordt hierbij vanuit gegaan dat dit uiteindelijk leidt tot een accurater and preciezer model, dan wanneer iedere data op zich wordt gebruikt. De uitdagingen van deze benadering zijn, voorspelbaar, bemonsteren en scoren: vele verschillende mogelijke modellen moeten gegenereerd (bemonstering) worden en geëvalueerd worden om de correcte oplossingen eruit te vissen (scoren). In dit proefschrift, focus ik voornamelijk op het incorporeren van cryo-electronenmicroscopie (cryo-EM) data en chemische kruisverbindingen gekoppeld met massaspectrometrie (CXMS) in het modelleringsproces. Cryo-EM is een snel ontwikkelende methode dat meestal lage-resolutie dichtheidsinformatie van grote macromoleculaire complexen oplevert, hoewel, met de huidige vooruitgang bijna-atomaire resolutie kan worden behaald voor specifieke systemen; CXMS daarentegen verschaft afstanden tussen residuen binnen het complex. Beide methodes leveren dus elkaar-aanvullende informatie op.

In **Hoofdstuk 1** geef ik een meer technische introductie tot integratief modelleren en de verschillende experimentele data die gebruikt worden. Ik introduceer de belangrijkste hedendaagse software paketten die een uitgebreid pallet aan integratieve methodes bieden voor macromoleculair modelleren, zoals het door data-aangedreven dockingsoftware HADDOCK, dat ontwikkeld is in het lab waar ik werk. Naast dit stel ik ook een nieuw concept voor, namelijk explorerend modelleren, waar de nadruk ligt op het kwantificeren en, indien mogelijk, visualizeren van de informatiehoeveelheid van de beschikbare data. De nadruk hier constrasteert met die van integratief modelleren, dat voornamelijk specifieke data-consistente structuren verschaft.

Hoofdstuk 2 beschrijft PowerFit, een hoge-prestatie software pakket en programma voor het automatisch plaatsen van hoge-resolutie starre biomoleculaire structuren in lage-resolutie cryo-EM dichtheidsmappen. PowerFit verricht een systematische zes-dimensionale zoektocht van de drie translationele en rotationele vrijheidsgraden om locale kruiscorrelatie-optima te bepalen voor het objectief plaatsen van structuren in cryo-EM dichtheden. Daarnaast introduceer ik een nieuwe en gevoeligere correlatiescore, de zogeheten kernverzwaarde locale kruiscorrelatie die de bruikbare resolutierijkwijdte verder uitbreidt voor het succesvol plaatsen. Po-

werFit is een eerste stap in dit proefschrift om hoge-resolutie structuurdata te combineren met lage-resolutie cryo-EM data.

Vervolgens in **Hoofdstuk 3** kwantificeer ik de resolutievereiste voor het succesvol plaatsen van starre biomoleculaire structuren in een cryo-EM map als functie van de grootte van het biomolecuul. Ook laat ik unambigu zien dat de kernverzwaarde Laplace locale kruiscorrelatie de best presterende score is. Als laatste, omdat de resolutievoorwaarde voor het succesvol plaatsen van een subeenheid vaak aanzienlijk lager is dan de resolutie van huidige cryo-EM data, kan dit uitgebuit worden door het multischaal afbeelding-pyramide concept, om de zes-dimensionale zoektocht significant te versnellen.

Hoofdstuk 4 bespreekt de implementatie van cryo-EM data in HADDOCK, wat resulteert in een waarlijke integratieve modelleringsbenadering, hetgeen de combinatie van alle andere HADDOCK-ondersteunde data toestaat. Het centrale concept hier is het gebruik van zwaartepunten, punten die het zwaartepunt van alle gedockte subeenheden voorstellen. HADDOCK laat het de gebruiker vrij om ambigue afstandsbeteugelingen te gebruiken, indien de locatie van verschillende eenheden niet kan worden onderscheiden. Het gebruik van de lage-resolutie cryo-EM data verhoogt aanzienlijk zowel het aantal alswel de kwaliteit van acceptable oplossingen in onze methode. Verder wordt het gebruik van deze krachtige integratieve methode gedemonstreerd op twee ribosoom en twee virus-antilichaam systemen, wat verdere inzichten van het tussenvlak oplevert en toekomstige experimenten kan leiden voor verder onderzoek.

De HADDOCK2.2 webserver wordt besproken in **Hoofdstuk 5**. De webserver biedt structuurbiologen een gebruiksvriendelijke interface tot het gebruik van onze geupgrade HADDOCK2.2 software. Noemenswaardige functies zijn de introductie van samengestelde molecuultypes, zoals eiwit-DNA complexen, en verdere op NMR-gebaseerde beteugelingen, zoals residuele dipolaire-koppelingen and vals-contactverschuivingen, wat nieuwe wegen opent voor macromoleculair docken. De webserver kan gratis gebruikt worden voor academische doeleinden via `http://haddock.science.uu.nl/services/HADDOCK2.2` na een simpele registratie.

Vanaf **Hoofdstuk 6** richt ik me op het gebruik van afstandsbeteugelingen in het algemeen en het gebruik van CXMS data specifiek. Ik introduceer nog een software pakket en programma, genaamd DisVis. DisVis

kwantificeert en visualizeert de informatiehoeveelheid van afstandsbeteugelingen via het concept van de toegankelijke interactieruimte, de telbare set van alle data-consistente oplossingen. Bovendien kan deze methode vals-positieve afstandsbeteugelingen identificeren, en het toont de zelfconsistentie van de data. DisVis is een eerste stap in het nieuwe concept van explorerend modelleren.

De methode wordt verder uitgebreid in **Hoofdstuk 7**, waar explorerend modelleren is gebruikt om tussenvlakresiduen te infereren. Ik toon aan dat tussenvlakresiduen kunnen worden voorspeld tot 90% accuraatheid voor starre bindingscomplexen als er 3 tot 7 langeafstandsbeteugelingen bekend zijn. De resulterende tussenvlakresiduen kunnen vervolgens gebruikt worden in HADDOCK als actieve en passive residuen om de robuustheid van de CXMS data te vergroten door dit te combineren met de standaard non-ambigue afstandsbeteugelingen.

In het laatste **Hoofdstuk 8** som ik mijn bevindingen op, en ik geef een persoonlijke kijk op het veld van integratief modelleren en stel verdere toekomstige onderzoeksvelden voor. Ik pleit voor het opleiden van de *hybride wetenschapper*, een expert op het tussenvlak van experimentele en computationele wetenschappen om het integratief modelleren verder te brengen. Verder bepleit ik een reïnterpretatie van structuurmodellen, aangezien zij niet beschouwd moeten worden as een enkele goedgedefinieerde structuur, maar eerder als een ensemble van modellen, zoals al eerder is voorgesteld in NMR spectroscopie. Vooral in integratief modelleren, waar de modellen meestal zijn gegenereerd met schaarse data, is de ensembleruimte groot en diffuus. Explorerend modelleren en het concept van de telbare toegankelijke interactieruimte kan helpen in de transitie naar de ensemble-interpretatie, en tegelijkertijd de informatiehoeveelheid van de data kwantificeren.

In het algemeen beschrijft dit proefschrift nieuwe benaderingen en computationele hulpmiddelen voor structuurbiologen om de experimentele data te interpreteren, dat van toenemende mate komt van diverse bronnen, en introduceert het een frisse kijk op modelleren via het concept van explorerend modelleren.

Acknowledgements

Now after 3 years 11 months and a couple days of hard work while finishing up the PhD and thesis, it's an apt time to reminisce about how I got here and give thanks where thanks is required and deserved. My scientific career starting really during my Bachelor in the Theoretical Chemistry group of **dr. Joop van Lenthe**, where I worked on some relativistic quantum chemistry in the GAMESS-UK package written in Fortran77. Even though the project was not really a success for several reasons, it fired up my interest into computational science, and taught me essential computer skills. Furthermore, Joop's approach to science and approachable demeanor are highly valued by me and I'm grateful for his supervision. Later in my Master I met high-potentials such as **Freddy Rabouw**, **Niek den Harder**, **Hinke Schokker** and **Marie Anne van de Haar**, together forming the *power trio*. Their approach to studying and making assignments was enlightening and made my life during the Master easier and filled with loads of fun.

Thanks to Niek, I decided to do my Master internship at the FOM Institute for Plasma Physics in Nieuwegein, where I eventually joined the MolDyn group under the supervision of **dr. Anouk Rijs**, and **Sander Jaeqx** (pronounce: Sjaaks). I had a blast for the year I was there, fully enjoying the work atmosphere, castle garden, excellent canteen, borrels, and great colleagues. It even resulted in my first (first-author) paper. So I'm very thankful to Anouk, who also gave an amazing speech during the Master ceremony. However, after my Masters I did want to return to full computational science and so I had to move on. After applying for a few positions that all were already filled and about to move on to an ordinary job, I came across an open position in the Computational Structural Biology group of **prof. dr. Alexandre Bonvin** and applied. Naturally, I was hired, and so my HADDOCKing time started ...

One of my first memories when I started out in the HADDOCK group is the presentation by **Ezgi Karaca** about SAXS scoring of protein models. The presentation contents and level really blew me away, and I was just so impressed by the whole group in general. Ezgi turned out to be a great role model for me during my PhD for several reasons: first of all, I'm the follow up "data-integrate-or"; furthermore, she published a Structure paper; she received a Keystone scholarship; and she's

just a great scientist and person in general. So I'm very much obliged to Ezgi for setting out a path to walk during my PhD. Another example was set by **Panagiotis Kastritis**. Panos' day typically started around noon and ended somewhere around midnight, a schedule that I also appreciate, and he showed incredible scientific hunger and interest, making him an inspiring fellow PhD student, whose desk I am now occupying after cleaning it up. On the same desk at the short end there sat **Mikaël Trellet**, a FIFA guru. Mikaël is one of the few people that know that I'm an amazing cook as I once made him my famous tagliatelle salmon-spinach-cream-fresh dish, and he was, according to an independent Greek source, the only guy who was funnier than me in the lab. The last person that initially sat in this office was **Marc van Dijk**, the only other Dutch HADDOCK person that I met during my PhD. So I want to thank him for being Dutch, for providing tips for creating an award winning poster, and initially helping me out with installing software on my Mac. **Christophe Schmitz** was another postdoc present in the beginning. Christophe has had a way bigger impact on my PhD than he probably realizes, as he referred within the HADDOCK CNS source code to the paper 'Quaternions in molecular modeling'. This paper has been indispensable in the creation of mainly DisVis and, to a lesser extent, PowerFit. So only for that I'm already very grateful. The senior postdoc of the lab, the one wielding the most power, was **Adrien Melquiond**, and I had the honour of him being my daily supervisor and co-promotor during the first part of my PhD. So thanks to Adrien for his supervision and helping me with brainstorming during the HADDOCK-EM project. The last person of what I call the 1st HADDOCK generation was **João Rodrigues**, my fellow PhD colleague for almost my whole stay and together with me the bridge to the 2nd HADDOCK generation. Known around the lab as the "Mini-BOSS", João was an unstoppable Powerplayer during CAPRI and organizing all kinds of stuff in and out the lab. He is also the one who came up with the name 'PowerFit' for the fitting software. So a huge thanks to him for making my stay pleasant and easier and even possibly delivering me a very nice postdoc position. After his departure I was the senior PhD of the group, and I felt that with great power there comes great responsibility.

The 2nd HADDOCK generation started with the introduction of some Italian blood from **Anna Vangone**. She once made a classical Italian dinner for the whole group, which was amazing. **Li Xue** was the next

person to join the group. I want to thank her for being the unscrambler of my amateurish Chinese speaking and her wise lessons of Chinese culture. The latest PhD addition to our group is **Cunliang Geng**, my current office mate and first PowerProtégé. I have seen Cunliang grow up since the start of his PhD to now, both in the lab, and in the gym, now almost being capable of doing a full free wide-grip chin-up. I am very proud of him and give him a big thanks for the interesting dinners in the lab. Finally, the newest member of the lab is **Zeynep Kurkcuoglu** further enforcing the girl power in the group. She is Docker-ifying PowerFit, a noble task, that I definitely appreciate.

This leaves of course the big cheese himself, the King of the Computational Structural Biology group **Alexandre Bonvin**. Now looking back at my interview I am a bit puzzled why I was hired, but it doesn't matter as the whole PhD experience has been an awesome ride for the last four years, and which worked out very smoothly. Thanks to Alex' leadership and guiding of the full group, working here just is a great experience with inspiring science. In an analogy to chess, Alex once emailed me that even though the King stay the King, he has limited moves and is depending on his army. All I can respond is that this pawn was very happy keeping the King in charge and is now about to promote. Overall I feel heavily indebted to him for providing me this opportunity and it has been an honour and a pleasure.

Next to the CSB group members there are of course also other people that I want to thank. For example people that I have shared an office with: **Klaartje Houben** and **Maryam Faridounnia** thanks for keeping me company the first years. Later on **Alma Svatoš** joined me in my own office. We had some interesting Wikipedia crawls and she was just a very funny office mate. She was followed up by **Siddarth Narasimhan**, who is now a fellow PhD student in the ssNMR group. Sid was often there also in the weekend, inhibiting me from playing The Police out loud. Still, it was fun to have him sitting at the short end of my desk, asking many questions.

I also want to thank the other people of the NMR lab in random order. **Mohammed Kaplan**, the man with many faces, who started and will graduate around the same time as I, and the initiator of many philosophical conversations. **Mark Daniels** for his always constructive remarks and creative impact during the writing of my papers and for

his excellent keepers training. **Eline Koers** for providing intense discussions and interesting opinions on diverse subjects. **Elwin van der Cruijsen** for setting up badminton and football teams and for keeping me off work. **Mehdi Nellen** who was a Master student and introduced me to the time management system of Pomodoro. **Ramon van den Bos** is another gym colleague with whom I also engage in public tennis. **Prof. dr. Marc Baldus** for the fun Italy road trip in the Fiat Cinquecento. **Markus Weingarth** for giving solid political advice, providing GPU resources, and his understandable enthusiasm for PowerFit. **Dr. Hans Wienk**, **prof. dr. Rolf Boelens**, **prof. emer. dr. Rob Kaptein**, **ing. Johan van der Zwan** and **dr. Gert Folkers** for excellent high-quality coffee table discussions. **Barbara Hendricx** for taking care of important bureaucratic arrangements. And the remaining PhD students **Deni Mance**, **Cecilia Pinto** and **Ivan Corbeski** for their company, and I wish them all the best in their coming PhD-years.

Publications

- **G.C.P van Zundert**, J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, E. Karaca, A.S.J. Melquiond, M. van Dijk, S.J. de Vries and A.M.J.J. Bonvin. The HADDOCK2.2 web server: User-friendly integrative modeling of biomolecular complexes. *J Mol Biol* Advanced Online Publication (2015).

- **G.C.P. van Zundert** and A.M.J.J. Bonvin. DisVis: quantifying and visualizing accessible interaction space of distance-restrained biomolecular complexes. *Bioinformatics*, Advanced Online Publication (2015).

- **G.C.P. van Zundert**, A.S.J. Melquiond and A.M.J.J. Bonvin. Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. *Structure* 23, 949–960 (2015).

- **G.C.P. van Zundert** and A.M.J.J. Bonvin. Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. *AIMS Biophysics*, 2, 73–87 (2015).

- M. Kaplan, A. Cukkemane, **G.C.P. van Zundert**, S. Narasimhan, M. Daniëls, D. Mance, G. Waksman, A.M.J.J. Bonvin, R. Fronzes, G.E. Folkers and M. Baldus. Probing a cell-embedded megadalton protein complex by DNP-supported soid-state NMR. *Nat Methods* 12, 649–652 (2015).

- **G.C.P. van Zundert** and A.M.J.J. Bonvin. Modelling protein-protein complexes using the HADDOCK web server. *Methods Mol Biol* 1137, 163–179 (2014).

- J.P.G.L.M. Rodrigues, A.S.J. Melquiond, E. Karaca, M. Trellet, M. Van Dijk, **G.C.P. Van Zundert**, C. Schmitz, S.J. de Vries, A. Bordogna, L. Bonati, P.L. Kastritis and A.M.J.J. Bonvin. Defining the limits of homology modelling in information-driven protein docking. *Proteins*, 81, 2119-2128 (2013).

- **G.C.P. van Zundert**, S. Jaeqx, G. Berden, J.M. Bakker, K. Kleinermanns, J. Oomens, A.M. Rijs. IR spectroscopy of isolated neutral and protonated adenine and 9-methyladenine. *ChemPhysChem* 12, 1921–1927 (2011).

About the author

Gydo Cornelis Petrus van Zundert was born in Oud Gastel on the 12th of June, 1987. He attended the local Sint Bernardus School, where he graduated with the maximum Cito score of 550 and co-starred in the final school play. After an uninspiring yet fun time at the VWO of the Sint Norbertus College in Roosendaal, Gydo started his Bachelor *Chemistry* in 2005 at Utrecht University. He performed his Bachelor thesis in the Theoretical Chemistry group of **dr. Joop van Lenthe**. Next, he registered for the Master *Nanomaterials: chemistry and physics*, and did his Master's internship in the MolDyn group at the FOM Institute for Plasma Physics Rijnhuizen, in Nieuwegein, under the tutelage of **dr. Anouk Rijs**. His work resulted in a first-author paper and he graduated with the maximum GPA of 4.0.

On the 1st of November 2011 he began his doctorate in the Computational Structural Biology group of **prof. dr. Alexandre Bonvin**. During these years he received a poster price (combined ACMM/NSBM symposium) and two travel grants (Keystone Symposia Future of Science Scholarship, and Instruct Biennial Meeting Scholarship), and he was the PhD representative of the NMR department in 2014 and 2015. His scientific work has been published in high-quality peer-reviewed journals. Gydo will defend his dissertation on the 25th of November 2015.