**Introduction**

**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 \cite[Nietzsche1891]. 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 \cite[Braun2012]. 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 \cite[Stein2011]. Inhibitors of these protein-protein interactions are an upcoming class of molecules with a profound impact on drug-development \cite[Wells2007].

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 \cite[Campbell2002] and a basis for rational drug design \cite[Bienstock2012, Sable2015]. 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 and flexibility of the macromolecules and weak, 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 for 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 \cite[Marti-Renom2000]. Macromolecular docking is the field that occupies itself with predicting the structure of a complex starting from their individual components \cite[Moreira2010], 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 \cite[Kundrotas2012]. However, this approach is only amenable for complexes for which co-crystallized

templates are available \cite[Vakser2013]. 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 \cite[Huang2015]; the data-driven approach tries to increase the success rate by including additional information from biophysical and biochemical methods during the docking \cite[Karaca2013, Rodrigues2014]. Data-driven docking is also more popularly known as hybrid or integrative modeling of biomolecular complexes.

**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 \cite[Alber2007a, Ward2013]. The procedure can be abstracted in four stages \cite[Schneidman-Duhovny2014]:

\startitemize[n]

\item Gathering information: collect information in the form of e.g. experimental data, bioinformatics predictions, statistical inference, just about anything that can be of use during the modeling.

\item Model representation and evaluation: the degrees of freedom of the model should be chosen, such as an all-atom model or a more coarse-grained representation, depending on the information content the data are providing. The data can be used to actively search for proper models and its information content should be transformed accordingly.

\item 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 3D rigid bodies, an exhaustive search might be performed, while for higher-dimensional systems Monte Carlo and simulated annealing approaches might be more efficient.

\item 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.

\stopitemize

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.

**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 about biomolecular complexes. NMR spectroscopy is also capable of pinpointing interface residues through the use of chemical shift perturbations (CSPs) \cite[Case2013], and the relative orientation of subunits to each other by residual dipolar couplings (RDCs) \cite[Chen2012], among several other methods \cite[vanIngen2014]. 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 shapes, though with low-resolution \cite[Kock2003, Putnam2007, Schneidman-Duhovny2012, Blanchet2013]. 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 \cite[Hopf2014]. 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.

**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 \cite[Milne2013]. In cryo-ET whole cell slices are studied by systematically tilting the sample and imaging projections from different angles; electron crystallography is mainly aimed at investigating membrane proteins that can form two-dimensional crystals or helices; 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 \cite[Glaeser1978]. To diminish this effect, the sample is typically plunge-frozen in liquid ethanol to instantly vitrify 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 combining the well-defined phases of the images with 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 \cite[Smith2014] to the rising star in structural biology (subtitle of the cryo-EM Gordon Research Conference 2014), to revolutionizing structural biology \cite[Bai2015, Nogales2015]. Although electron diffraction resolution has remained the same at around 2Å \cite[Gonen2005], cryo-ET's subtomogram averaging now attains sub-nanometer resolution \cite[Schur2013], and the single-particle cryo-EM resolution record for now stands at 2.2Å \cite[Bartesaghi2015].

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 10Å, secondary structure elements at 9Å, the separation of beta-sheets and bulky side-chains at around 4Å \cite[Baker2010]. 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.

**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 \cite[Tran2015]: 1) the cross-links are added to the sample after optimizing the reaction conditions, and 2) the cross-linked proteins are isolated to reduce the number of false-positives; 3) is to digest the cross-linked proteins into peptides using trypsine or other proteases, after which 4) the peptides are enriched using physico-chemical methods, such as size exclusion chromatography, affinity chromatography, 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 \cite[Leitner2010, Merkley2013]. However, the major bottleneck is the final data analysis as millions to billions of fragments can be produced and need to be considered \cite[Tran2015]. 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 \cite[Tosi2013], the Polycomb Repressive Complex 2 \cite[Ciferri2012], and the 30S-elF1-elF3 translation initiation complex \cite[Erzberger2014].

**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 \cite[Leaver-Fay2011], the second is the Integrative Modeling Platform (IMP) developed by the Sali lab \cite[Russel2012], and the third is our in-house data-driven docking software HADDOCK (High Ambiguity Driven DOCKing) \cite[Dominguez2003, deVries2007].

*Rosetta*

Rosetta is at its core a structure prediction software package, and is well known for its elaborate and accurate force field and conformational sampling techniques \cite[Rohl2004]. Although originally a de novo protein prediction program \cite[Simons1999], it has ventured into a more integrative approach and can now also perform X-ray crystallography refinement (MR-Rosetta) \cite[DiMaio2011], use NMR data (CS-Rosetta), CXMS data \cite[Kahraman2013], and recently also cryo-EM data \cite[Demers2014, DiMaio2015], resulting in the current prediction software package juggernaut that it is today \cite[Adams2013]. The Rosetta source code was recently rewritten with the release of Rosetta3 \cite[Leaver-Fay2011]. 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 \cite[Alber2007, Alber2007a]. The IMP software consists of several user interface layers, each giving more control to the user \cite[Russel2012, Webb2011]. 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 \cite[Lasker2009, Lasker2010] and FoXS for the calculation of SAXS curves \cite[Schneidman-Duhovny2010]. In addition, the IMP package is also integrated into the molecular graphics visualization program UCSF Chimera \cite[Yang2012]. 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 \cite[Dominguez2003]. Since then, HADDOCK's capabilities have steadily increased, and now also supports the use of RDCs \cite[vanDijk2005], relaxation anisotropy \cite[vanDijk2006], protein-DNA docking \cite[vanDijk2006], solvated docking using explicit water \cite[vanDijk2006a, Kastritis2012], docking up to 6 components \cite[Karaca2010], NMR pseudocontact shifts \cite[Schmitz2011], SAXS and collision cross sections derived from MS \cite[Karaca2013], and protein-peptide docking \cite[Trellet2013]. The HADDOCK webserver was introduced in 2010 \cite[deVries2010] 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 backend \cite[Brunger2007].

**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, which does not necessarily provide insight into the information content of the restraints. 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 the model uncertainty, and guide future work.

**Overview of thesis**

This thesis primarily describes new computational methods to handle cryo-EM and distance restraints data for integrative and explorative modeling. In \inchapter[chapter:powerfit] 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 \inchapter[chapter:image-pyramids] 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. \inchapter[chapter:haddock-em] 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 \inchapter[chapter:haddock2.2-webserver] I present the HADDOCK2.2 webserver, an upgrade of the HADDOCK webserver, for user-friendly integrative modeling of biomolecular complexes. \inchapter[chapter:disvis] deals with quantifying and visualizing the information content of distance restraints in general, and cross-link 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 \inchapter[chapter:inferring-interface-residues], 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, I present a summary of the thesis and provide a personal perspective on the field of integrative modeling, proposing further lines of research.