Computational structural biology of macromolecular interactions

João Pedro Garcia Lopes Maia Rodrigues

ISBN/EAN xxx-x-xx-xxxxxx-x
Doctoral thesis
Computational Structural Biology of Macromolecular Interactions
João Rodrigues
NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Faculty of Chemistry,
Utrecht University, The Netherlands
December 2014

Printed in The Netherlands by Ridderprint BV

Computational Structural Biology of Macromolecular Interactions

Computationele structuurbiologie van macromoleculaire interacties
(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 vrijdag 12 december 2014 des ochtends te 10.30 uur

door

João Pedro Garcia Lopes Maia Rodrigues

geboren op 29 juli 1986,
te Coïmbra, Portugal

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

Beoordelingscommissie:

Prof. dr. Rolf Boelens
Prof. dr. Antoinette Killian
Prof dr. Michael Levitt
dr. Ed Moret
Prof. dr. Berend Snel

Dedicado à minha família e amigos
Obrigado por me empurrarem para esta aventura

Table of Contents

General Introduction

"Research is formalized curiosity. It is poking and prying with a purpose."

- Zora Neale Hurston

Curiosity is a natural trait of human beings. We are born curious, constantly exploring our surroundings, eager to learn everything about our environment. And as we grow, and develop the ability of abstract thought, we gaze inwards, towards ourselves, and start asking questions about our condition as human beings, about our own bodies, and ultimately, about how we came to be. This very curiosity led Mankind to invent, or rather discover, science, and, as knowledge was amassed, to subdivide it in disciplines, or fields, such as biology and chemistry.

The term *biology* stems from the Greek word βίος (*bios*, life) and the suffix λογία (*logia*, study of). Despite being a relatively novel term – it was first used in Latin form by Linnaeus in 1736 – it relates to a scientific discipline far older. Ancient civilizations such as the Greeks and Egyptians are known for their inquisitions on the functioning of the human body and on the natural order of life in general. The circulatory system, for example, was described as early as 1600BC. However, these ancient scientists were limited by their physical senses. They could only describe what they could see or feel. The invention of the microscope, disputed by several parties, lifted this barrier and allowed observation of what Robert Hooke called 'minute bodies'. In his book *Micrographia*, published in 1665, Hooke showed the world the microscopic details of a fly's eye and perhaps more importantly, of a walled component of plants: a cell. Only a few years later, Antonie van Leeuwenhoek would communicate to the Royal Society illustrations of other microorganisms, neither plant or animal. These and other groundbreaking observations of the microscopic world changed the way of biological sciences, urging the revision of old theories on the origin of life and opening a whole new field of study: cellular biology.

The progress made in the biological sciences was mirrored by other disciplines, such as chemistry. Unlike the noble principles that guided biological research, chemistry was for a long period driven by greed. Alchemy, from where the term chemistry is thought to derive, was a mixture of chemistry, astronomy, mysticism, and metallurgy whose practitioners – the alchemists – sought to master to produce gold and the elixir of eternal life. Nevertheless, there were those who attempted to define general principles about matter. The Greek philosopher Aristotle, without knowledge of atoms or molecules, postulated a theory according to which all matter was derived from a combination of five elements: water, fire, earth, air, and ether. The leap from philosophy to a hypothesis-driven discipline happened, as with biology, during the mid-1600s, when Robert Boyle published his text *The Sceptical Chymist*, which refuted Aristotle's theories – still held in high regard - and hypothesized that matter is made of atoms and that their motion and collisions were responsible for all phenomena. This separation of chemistry from the mystical alchemy marked the foundation of modern chemistry and ignited a revolution that we are still, somewhat, experiencing.

As the scientific revolution touched every realm of human knowledge it made us realize that perhaps the several disciplines were not divergent as we once thought. As biology advanced, scientists started realizing that all organisms – animals, plants, fungi, bacteria – were built from the same components – cells. As chemistry advanced, elements were discovered and cataloged, their properties were measured and recorded, and soon it was understood that we too were made of those same elements. Biology and chemistry, once so different, began to converge on the molecular level. The first step towards the merger was perhaps the synthesis of urea, until then thought to be unique to living organisms, by Friedrich Wöhler in 1828. This seemingly simple experiment dazed the scientific community, who believed in a stark separation of 'living' (organic) and 'dead' (inorganic) matter. Another significant contribution was the accidental discovery of Eduard Büchner, in 1897, that the extract of yeast cells was enough to catalyze the fermentation of sugar to alcohol. Both these discoveries, and many others in the decades in-between, cemented the view that what was

once viewed as 'life' was but a strictly regulated set of chemical reactions happening in a controlled environment – the cell – and carried out by very specialized molecules.

By the mid-1800s, the question was not *if* but *which* molecules populated the cell. Proteins were the first class of cellular molecules to be identified and studied. Antoine Fourcroy, a contemporary of Lavoisier, isolated vegetable proteins and observed that their properties were very similar to egg albumin. In 1838, the Dutch chemist Gerardus Mulder decomposed vegetable and animal proteins into their atomic elements and realized they all shared the same composition. Further, he proposed that proteins, being much larger than the molecules studied then by chemists, were composed of a unitary substance he called *grundstoff*, and proved it using acid hydrolysis to decompose proteins into what we now know to be amino acids. By 1902, with most natural amino acids already identified, Emil Fischer was able to synthesize peptides – protein fragments - in laboratory conditions and, simultaneously with Franz Hofmeister, propose a chemical explanation for the peptide bond. Therefore, by the first decade of the 20th century, biological chemists knew what proteins were, what they were made of, and how to produce and isolate them. They had yet to find, however, their true role inside the cell.

Moritz Traube was a good example of curiosity. Despite having never been associated with a university or research institute, Traube, a German wine merchant, published a treaty in 1858 on the study of fermentation from a purely chemical perspective. He established enzymes as protein-like compounds that required direct molecular contact with their substrate. He also categorized them by their function, much like we do nowadays. It is crucial to realize that this happened 40 years before Buchner's discovery of cell-free fermentation and 20 before Wilhelm Kühne created the word enzyme. Also Emil Fischer, in 1894, proposed what is still held today as a hallmark model of enzymatic mechanisms: "*To use an image, I will say that an enzyme and a glucoside [i.e., glucose derivative] must fit like a* ***lock and key****, in order to be able to exert a chemical effect on each other*". Three years later, between 1896 and 1897, the rector of Utrecht University Cornelius Pekelharing obtained highly active pepsin (an enzyme) preparations and recorded their positive reaction to protein tests. He therefore suggested that enzymes might be proteins. This opinion was not against the community; in fact, drawing on the influence and knowledge from inorganic chemistry, scientists believed that proteins were yet another molecule in a colloidal solution that composed enzymes. The centerpiece of the enzyme, however, was thought to be an inorganic element, which proteins helped carry. The realization that enzymes and proteins were strictly the same came only much later, well into the 20th century. In 1926, James Sumner crystallized the enzyme urease from jack bean and analyzed it, proving that the pure compound was a protein. The discovery was met with skepticism and Sumner heavily criticized by his peers. Richard Willstätter, the 1915 recipient of the Nobel Prize in Chemistry, had been working on enzyme catalysis since the early 1920s and had produced several preparations of pure horseradish peroxidase that showed no reaction for normal protein tests. Little did he know however, that the extreme dilution of his preparations was the cause behind the seemingly negative tests. Following Sumner's footsteps, John Northrop (pepsin, 1929) and Moses Kunitz (trypsin, 1932) provided enough evidence to turn the tide and establish the nature of enzymes as proteins. Both Sumner and Northrop would eventually see their efforts recognized by the scientific community, receiving the Nobel Prize in Chemistry in 1946.

It took only one century for biology and chemistry to coalesce and produce enough knowledge to provide a unified framework that explained the molecular foundations of the cell. Living cells are made of molecules such as proteins and nucleic acids, which respect the general rules of chemistry. They are synthesized from elementary components, interact with other molecules, and can be degraded again into components. Alongside the structural and

chemical properties of these molecules, scientists took interest on the nature of interactions in which the molecules participated. Fischer proposed the lock-and-key model for enzyme/substrate interactions in 1894. Leonor Michaelis and Maude Menten, drawing on the work of Victor Henri, established the mathematics of enzyme kinetics in 1913. Throughout the first half of the 20th century, these interactions were put in a cellular context by the work of Hans Krebs, Fritz Lipmann, Carl and Gerty Cori, Hans Kornberg, and others, who revealed the existence of highly regulated pathways of molecular interactions in the cell. These were marvelous bodies of work, especially if one considers that these researchers had as much insight on the systems they were studying as medieval scientists on microbes. Despite the microscope, and despite all the technological advances of the early 19th century – x-ray crystallography, quantum physics, and quantum chemistry – technology was just not enough to allow direct observation of molecular systems such as proteins and their interactions. Consider the following metaphor: imagine an observer atop a very tall building; if he were to look down at the streets, he could distinguish people from cars, and cars from buses, perhaps isolate particular individuals from a crowd. Given enough time, he would start to understand patterns such as the direction of the traffic, traffic lights, and even regular bus routes. Yet, the cornerstone of that society, human interaction, would be beyond his reach, invisible from such a distance. He could only propose theories on how that society worked based on the coarse observations he had made. This was the state of molecular biochemistry before the 1950s: carried out with scarce knowledge of the molecular world.

In 1912, a father and son team, William Henry Bragg and William Lawrence Bragg, proposed the theoretical rules of x-ray diffraction and thus enabled the imaging of molecules at atomic resolution. Molecules are shot with a beam of x-rays, a diffraction pattern is recorded, and a mathematical treatment converts this pattern into an electron density map from which atomic positions can be distinguished. In short, the method photographs molecules with enough resolution to see individual atoms. In the beginning, the systems that could be imaged using this technique were small. The first atomic resolution structure was that of table salt, sodium chloride, in 1914. The structure of diamond followed soon after. After two decades of developments, structures of biologically relevant molecules started appearing, mostly by Dorothy Hodgkin: cholesterol (1937), penicillin (1946), and vitamin B12 (1956). Meanwhile, the younger Bragg settled in Cambridge, England, as the head of the Cavendish Laboratory. Under his aegis, in February 1953, James Watson and Francis Crick produced in Cambridge the first three-dimensional model of a biological macromolecule – the DNA double helix. A couple of years later, John Kendrew and Max Perutz would replicate the feat, solving the three-dimensional structures of myoglobin and hemoglobin, respectively. All these discoveries were almost immediately recognized for their immense importance to molecular biology, earning Watson, Crick, Kendrew, and Perutz, Nobel Prizes in Physiology and Medicine and Chemistry.

As the methods to crystallize and analyze the diffraction patterns of x-ray experiments advanced, more and larger systems became amenable to this technique. In the fifty years that followed these discoveries, x-ray crystallography helped solve the structure of nearly a hundred thousand proteins and other biological molecules, offering a clear picture of how these molecules look like. Many of these structures portray interactions: proteins with other proteins, proteins with nucleic acids, and proteins with ligands. These interactions also gained relevance in the overall picture of cellular metabolism. Proteins not only accelerated chemical reactions via enzymatic catalysis, they were shown to be actively interacting with each other, transmitting signals, carrying other molecules, participating in the maintenance and replication of the genetic material, etc. X-ray crystallography also contributed to image these interactions and validate the theories proposed decades earlier (e.g. lock-and-key): in

1973, Robert Huber and co-workers published, in the Journal of Molecular Biology, the first crystal structure of interacting proteins, that of bovine trypsin and its inhibitor (BPTI).

Adding to the metaphor of the observer, imagine he is now armed with a very powerful photography lens and an unlimited number of storage cartridges. He can now zoom in on the ground level and focus on individuals, discern facial expressions such as a smile or a frown, or photograph a particular encounter or group of people and later analyze who was communicating with whom. The question is, though, is this enough to grasp the dynamics of society? As we know, a picture gives only a static glimpse of a moment. For most purposes, it suffices, and so do crystallographic structures. Yet, sometimes, there is the need for a continuous representation that illustrates motion. One can argue that several pictures can be taken in rapid succession and then collated to produce such a dynamic representation. After all, it did settle the question of Leland Stanford on whether a galloping horse ever lifts all four feet completely off the ground - it does. But what if this motion is too fast? What if it is so brief, that despite a very powerful lens firing shots repeatedly, it sneaks between two frames leaving the false impression of stasis? In x-ray crystallography of biological molecules, motion is a known problem, and often leaves a mark in the solved structure. Just as a photograph can become blurry if the shutter speed is not high enough, the electron density of highly dynamical regions of molecular structures is less clear and may impair the determination of the atomic positions. Perhaps more importantly, when studying systems such as protein-protein complexes whose interactions can be weak and short-lived, crystallization and proper diffraction may be impossible altogether.

The solution to this problem lies then in developing alternative methods that can monitor molecular motions in real-time, with atomic resolution. While it is not currently possible to directly image molecules, in real-time, in their native conditions (i.e. the cell), and at the atomic scale, it is possible to do so indirectly. Imagine giving our observer a microphone so he can eavesdrop on communications between people. He cannot directly observe interactions between people, but he can listen to their conversations and, given enough time, characterize a particular person with very high accuracy: their habits, their routine, their friends. The molecular counterpart of this 'microphone' is rooted in quantum physics and chemistry: nuclear magnetic resonance spectroscopy. Developed in the 1940s, it is a method akin to tuning into the frequency of particular atomic nuclei. Each individual nucleus – carbon, hydrogen, nitrogen, etc. - reports at a frequency that depends on its chemical environment and can therefore identify the atom (often) unambiguously. By recording the radio waves over time, since the molecules are not crystallized but free in solution, the spectrometer obtains a series of signals that have a time dependency and relate to the natural motion of the molecules. Mathematical post-processing can derive distances and angles between the nuclei, and thus pinpoint the relative location of each atom at a particular time-frame, allowing the reconstruction of a molecular structure. Beyond structure, nuclear magnetic resonance spectroscopy can also be used to gather information on the molecular motion of a particular protein for example, or to capture the dynamics of its interaction with a partner, via relaxation rates. Unfortunately, much like a radio, too many atoms broadcasting in one frequency will confound the listener. Although small proteins and their complexes are perfectly well suited for such analysis, large molecular systems are not. Furthermore, to be accurate, nuclear magnetic resonance spectroscopy requires isotope-labeled samples, which might become extremely expensive for very large systems that require several rounds of experiments. Other experimental methods have since been developed to explore systems not amenable to x-ray crystallography or nuclear magnetic resonance. Low-resolution cryo-electron microscopy and small angle (x-ray/neutron) scattering were developed and applied to large biological systems in recent years, and when combined with the previous

techniques, can produce insightful structural models. Nonetheless, the limitations, material expenses and human resources required by each one of these techniques, in particular when applied to extremely challenging projects with few guarantees of success, begged for the development of alternative, cheaper, and complementary methods.

In 1967, around the same time protein x-ray crystallography was booming, John Kendrew suggested to a young student a visit to the Weizmann Institute in Israel, in order to work on computational methods to simulate molecular systems. Computers had been invented only a couple of decades before, and their application to physical problems, namely many-body problems, had been immediate. From physics to chemistry, and then to biology, was a small step that required only enough computer memory to be available. These requirements were met in the late 1960s. The Weizmann Institute had built a computer in 1963 and affectionately named it the *Golem*. Shneior Lifson, then the institute's Scientific Director, had proposed a generic method to simulate any molecular system *in silico* using a simple mathematical function describing its potential energy – the Consistent Force Field. When Michael Levitt flew from Cambridge to Israel in October 1967, he started to work together with Arieh Warshel, a doctoral student of Lifson, on building a computer program that could calculate the energy and forces of any molecular system. After a few months, the program was complete, and while Warshel used it for small molecules, Levitt applied it to myoglobin and lysozyme. The outcome of this work was the first energy minimization (or refinement) of a protein structure. For those outside the field of protein structure determination, it is hard to grasp the impact of this publication. Watson and Crick's model of DNA had been built *manually* and so had Kendrew's myoglobin and Perutz's hemoglobin models. These were molecules with hundreds of atoms, each of them carefully interpreted from the electron density map and mapped into three-dimensional coordinates in a wire-frame model. These models also had to respect what was known from the chemical properties of the atoms: bond lengths, bond angles, all these properties were documented, after years of experiments, and had been determined to be quite constant. Levitt and Lisfon's computational approach offered a tremendous help: it combined all these definitions of lengths and angles and even long-range forces such as electrostatics – not used in the original paper – and van der Waals into a 'force field', which could be applied iteratively to an existing model in order to bring it to a minimum of potential energy. In other words, if the modelers had made a mistake and made a particular bond a bit too long or an angle too tight, the program could correct these and return the ideal coordinates of the molecule, as encoded in the force field.

A decade later, in the summer of 1976, Herman Berendsen, then a professor at Groningen University, organized an eight-week workshop at the CECAM (Centre Européen de Calcul Atomique et Moléculaire) center in Orsay, France. The workshop was titled 'Models for Proteins' and its goal was to bring together physicists and biologists to work on the then booming field of computational biology. Among the attendees were Michael Levitt, Martin Karplus, Wilfred van Gunsteren, Andrew McCammon, and others who would later become the leaders of the field. During those eight-weeks, intense collaboration and discussion led to several important developments, the most notable being the first molecular dynamics simulation of a biological molecule (the protein BPTI) by McCammon and Karplus. One year later, Shoshana Wodak and Joel Janin published the first protein-docking algorithm, based on Levitt's work, which enabled the study of molecular interactions *in silico*. Altogether, these efforts showed that the study of molecules was not limited to experiment; they could be modeled theoretically, in a computer, and the results were reasonable enough to aid experimental research.

The advent of computational structural biology, the branch of biology that concerns the study of the three-dimensional structure of biological molecules using computers,

had a significant and immediate impact on the scientific community. As years passed, computers became more powerful, and better methods to simulate proteins and other cellular inhabitants were developed and applied to several domains and several problems. These methods allowed scientists to finally surpass the resolution limit and represent, in real-time and at atomic resolution, pretty much any molecular system existing in the cell. In addition, instead of being considered adversaries or competitors, computational methods were recruited by many experimental techniques to help solve their own shortcomings. They now permeate every structural biology method used in science, from x-ray crystallography to nuclear magnetic resonance spectroscopy. In fact, it is perhaps because of computation that these experimental methods have become so widely used. The development of novel pharmaceutical drugs is nowadays preceded by computational analysis. Disease-causing mutations at the DNA-level are rationalized using protein structure prediction and dynamics. Metabolic pathways and their individual steps can be optimized or altered – *designed* – to suit the need of the researcher. But as much as its pioneers would have wished for, computation was not a panacea. It too suffers from its own shortcomings, often owing to limitations in the raw power provided by computers and by inaccuracies present in the force fields governing the simulations.

Despite all these centuries of research and all the advances, there are fundamental chemical questions that still lack an answer. The solutions proposed by the 18th and 19th century chemists were appropriate for them because they could not observe their systems in appropriate detail. The quantum revolution of the 1930s showed that, as we approach the atomic scale, laws such as those that Coulomb once formulated are not *so* accurate anymore. However, we cannot include the more accurate quantum laws in our simulations and indiscriminately apply it to biological systems; our current hardware is just not powerful enough to return meaningful results in a reasonable period of time. Classical mechanics, as pioneered by Newton, offers still the best trade-off in speed vs. accuracy. Another issue concerns the choice of systems under study. To simulate extremely large and dynamical systems such as the nuclear and cell membranes with atomic detail requires substantial computing power, which even the most powerful supercomputers in the world cannot provide. Consequently, on the verge of half a century of rapid developments, the field of computational structural biology requires constant attention and innovation in order to remain useful to the broader scientific community tackling ever more challenging problems.

This thesis describes several advances in the field of computational structural biology of macromolecular interactions, relying on a combination of new algorithms and the integration of both experimental and computational methods to yield better structural models of protein interactions. These advances are presented in five chapters:

The ***first chapter*** addresses the shortcomings of conventional *ab initio* protein docking algorithms and elaborates on a class of algorithms that can use any sort of information on the interaction to bias the numerical simulations: integrative modeling.

The ***second chapter*** assesses the usage of computationally derived models, based on the premise of homology, and sets the limits at which these models stop being reliable in integrative modeling calculations.

The ***third chapter*** introduces and evaluates co-evolution analysis methods for the prediction of biological interfaces between interacting proteins, an alternative to when experimental methods cannot produce (enough) information to guide the energy calculations and simulations.

The ***fourth chapter*** details the re-design and optimization of HADDOCK, the pioneer of integrative modeling of biological complexes, including the implementation of a coarse-grained force field and the expansion to n-body systems.

The ***fifth and last chapter*** introduces a new structure clustering method, based on a new distance measure, the fraction of common contacts, and an adapted asymmetric clustering algorithm, that has applications in large-scale studies and multiple-component and symmetrical systems.

The thesis ends with a short perspective on the current state-of-the-art and future challenges of the field of computational structural biology of macromolecular interactions.

Chapter 1

Integrative Computational Modeling of Protein Interactions

J. Rodrigues and A.M.J.J Bonvin

Published in 2014 in the FEBS Journal, volume 281, issue 8

Abstract

Protein interactions define the homeostatic state of the cell. Our ability to understand these interactions and their role in both health and disease is tied to our knowledge of the three-dimensional atomic structure of the interacting partners and their complexes. Despite advances in experimental structure determination methods, the majority of known protein interactions are still missing an atomic structure. High-resolution methods such as x-ray crystallography and nuclear magnetic resonance spectroscopy struggle with the high-throughput demand, while low-resolution techniques such as cryo-electron microscopy or small angle x-ray scattering provide too coarse data. Computational structure prediction of protein complexes, or docking, was first developed to complement experimental research and has since blossomed into an independent and lively research field. Its most successful products are hybrid approaches that combine powerful algorithms with experimental data from various sources to generate high-resolution models of protein complexes. This mini-review introduces the concept of docking and docking with the help of experimental data, compares and contrasts the available integrative docking methods, and provides a guide for the experimental researcher for what types of data and which particular software can be used to model a protein complex.

Introduction

The cell is a busy environment: several thousand molecules constantly meet and exchange information to define its metabolic state. Of all contributors to cellular homeostasis, proteins are the most abundant and active. Therefore, eavesdropping on their interactions and learning how information is being shared is pivotal for a complete understanding of the cell. Further, this also provides the first step towards rational development of therapeutics for many deadly or incapacitating diseases[1].

Cellular and molecular biology has evolved and delivered powerful methods to identify and pinpoint the cellular location of proteins and their interactions. Yet, only structural biology can provide definite answers on the mechanisms of these interactions by revealing the high-resolution atomistic structures of the underlying biomolecular complexes[2]. Experimental structure determination of interactions can be, however, a laborious, time-consuming, and costly endeavor[3]. The growing gap between the universe of known sequences and that of determined structures is enough evidence that high-throughput structural biology is a dream yet to come true[4,5]. This gap widens when considering the number of available structures of biomolecular complexes.[6] Computational structural biology, on the other hand, has the potential to deliver (high-resolution) models of protein-protein interactions. It has however struggled with inaccuracies since its inception in the late 1960s[7]. As a result, models have often been met with skepticism. Fortunately, the last decades have seen fascinating developments both in software and hardware[8] and computational structure prediction is nowadays routinely considered an integral part of research.

The docking field, in particular, has thrived in the last decade since the inception of the CAPRI (Critical Assessment of PRedicted Interactions) experiment[9]. Several rounds of blind predictions have stimulated discussions and resulted in the development of a wide variety of docking algorithms, some of which have consistently met considerable success[10-13]. Interestingly, the most successful participants of the latest CAPRI rounds fall in the category of "data-driven", "information-driven", or "integrative modeling" algorithms[14]. These approaches were developed to counter the inaccuracies of computational sampling and scoring methods by feeding them whatever experimental information is available for a given interaction. As recent CAPRI assessments have shown, this synergy is often enough to drive the docking calculations toward the right answer. As such, computational modeling of complexes has grown into a well-accepted method complementary to classical experimental techniques such as nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography.

This mini-review is dedicated to the docking approaches for the modeling of biomolecular complexes, with a focus on integrative and information-driven algorithms. It first explores what computational docking is and what are its limitations, then explains which types of information docking methods best benefit from and why including information is a valuable strategy. Alternative methods to predict the structure of protein-protein complexes are then shortly discussed. It closes with an overview of the current and future challenges for the docking community.

Computational molecular docking

Molecular docking is usually defined as the structure prediction of a molecular complex starting from the individual structures of its participants. Despite their differences, docking algorithms share three common elements: first, three-dimensional (3D) structural models of

the individual components must be available, whether experimentally determined or computationally predicted; second, they must explore the conformational landscape of the interaction and generate structural models of the complex, what is called sampling; finally, they must assess the generated models and select those that are more likely to be representatives of the native complex, what is called scoring. 3D structures of the components are usually obtained experimentally via X-ray crystallography or NMR spectroscopy. Nevertheless, considering that the number of known protein sequences dwarfs the number of available 3D structures, it is rather common to obtain a model of the protein in question through homology modeling methods[6,15]. As for sampling and scoring, these are the "Achilles heel" of docking and remain two very difficult problems given the natural properties of biomolecules, in particular their flexibility, and our limited understanding of their thermodynamics[16-18].

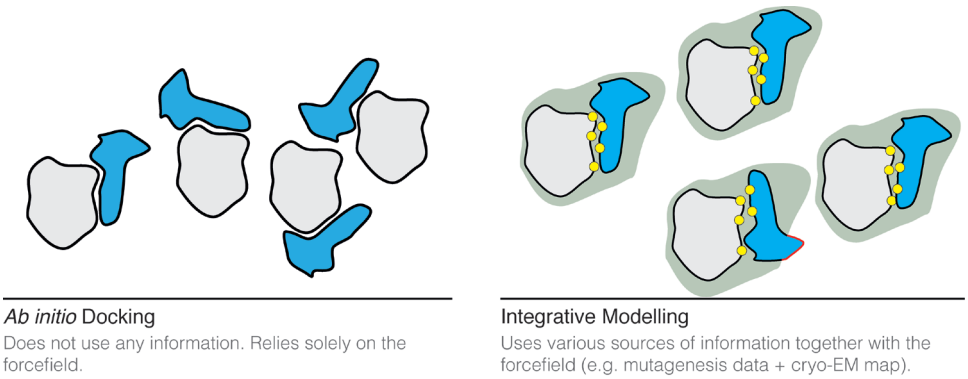


Figure 1. *Ab initio* vs. Integrative Docking. Docking algorithms can use information extracted from experiments or (bioinformatics) predictions to narrow down the possible orientations of the complex. By adding external information, the sampling can be limited to the regions defined in the data and the number of wrong solutions can be reduced. Multiple sources of information can be used to solve ambiguities and improve the convergence of the calculations to a few conformations. Nevertheless, the inclusion of experimental/prediction data does not guarantee that all solutions will be correct.

Sampling: exploring the conformational and interaction landscape

Sampling, simply put, is the enumeration of all possible orientations and conformations the monomers of a complex can assume in 3D space. If we assume both molecules as rigid, i.e. that their conformations pre- and post-binding do not differ substantially (~< 1Å root mean square deviation of the coordinates of the backbone atoms), then there are methods that can efficiently cover the entire search space[19]. These are usually based on grid-searches using fast Fourier transforms (FFT), as pioneered by Katchalski-Katzir *et al.*[20] and widely used nowadays[21-23], geometric hashing, as first developed and still used by Wolfson and colleagues[24,25], or spherical harmonics, as introduced and used by Ritchie in HEX[26,27].

FFT techniques represent the proteins" surfaces in a Cartesian grid model that favors close contacts, i.e. overlap of the surfaces, and penalizes overlap of the core, performing exhaustive rigid-body conformational searches very efficiently (typically in a few hours). Geometric hashing techniques divide the molecular surface in interaction patches and

match them across the interacting molecules. Spherical harmonics-based methods also calculate Fast Fourier correlations but are computationally more efficient due to the usage of a combination of spherical harmonic functions to describe the protein shapes and the calculation of docking orientations via scalar products of rotated and translated coefficient vectors. All these methods have been designed to evaluate molecular shape complementarity but often incorporate simple energy functions to bias the scoring, for example based on electrostatics (e.g. ZDOCK[21] and PyDOCK[23]), desolvation (e.g. pyDOCK) or statistical potentials (PIPER[28], IRAD[29]). Other methods are less computationally efficient but, nevertheless, powerful. HADDOCK[30] uses a derivative-based search method in Cartesian space – rigid-body energy minimization – that acts directly on an energy function represented by a sum of electrostatics, van der Waals, and restraint energies, therefore targeting specific patches on the molecular surface deemed favorable by the energy function. A relatively recent method, SwarmDock[31], incorporates normal-mode analysis – an approach pioneered by ATTRACT[32] - into a Particle Swarm Optimization meta-heuristic, effectively docking while optimizing conformation, position, and orientation simultaneously.

Unfortunately, proteins tend not to be rigid[33] and their flexibility causes challenges for docking[16,34,35]. Sampling the full conformational landscape of a biomolecule is a well-known problem in protein folding: it is time-consuming, suffers from inaccuracies related to the energy functions used to describe the landscape, and usually requires an army of computing CPU cores or specialized hardware[8]. Therefore, most docking methods refrain from using traditional molecular dynamics simulations in Cartesian space to sample the flexibility of the interacting partners upon binding. Instead, they often implement other (more efficient) search methods such as Monte Carlo, (e.g. RosettaDock[36]), normal-mode analysis (e.g. ATTRACT[37], SwarmDock and FiberDock/FlexDock[25]), simulated annealing in torsion angle space (e.g. HADDOCK), and/or use simplified representations of the system such as coarse-grained models (e.g. ATTRACT[38] and RosettaDock). Finally, some methods have been developed to tackle the particular problem of dealing with large domain motions, such as the flexible multi-domain approach of Karaca & Bonvin that can describe extremely large conformational changes (up to 19.5Å RMSD between the free and the bound state)[16].

Molecular dynamics in Cartesian space is still often used to refine the docking models, in some cases in explicit solvent to better represent the cellular environment, either as last step in the docking protocol (e.g. HADDOCK) or as an additional step (e.g. ATTRACT[38]).

Scoring: discriminating right from wrong

The so-called "holy grail" of docking is to develop a method – scoring function - that is able not only to discriminate near-native conformations from others, but also to accurately estimate the binding affinity of the interacting molecules[17,39]. Quantum mechanical descriptions of the molecules and the interaction medium (e.g. aqueous solution) could provide such discrimination but are too computationally demanding to be applied to protein-protein complexes, let alone the thousands of models produced by docking methods. Consequently, docking programs implement scoring schemes based on simpler molecular mechanics, empirical observations, evolution/homology, or a combination of these[39]. Furthermore, scoring functions tend to be adapted to particular sampling schemes; i.e. sampling methods that generate a very large number of models (e.g. FFT-based, geometric hashing, rigid-body docking) tend to use simple but fast-to-compute scoring functions, while methods that generate a smaller number of models (e.g. simulated annealing, molecular dynamics, Monte

Carlo) can afford to employ more elaborate, and therefore slower, scoring functions. In line with this, different stages of the same docking approach often use different scoring functions, typically increasing the complexity (and CPU costs) of the scoring function used while reducing the number of models considered.

Electrostatics interactions are commonly included in scoring functions through a Coulomb potential, even in fast, exhaustive methods such as FFT-based Hex, ZDOCK, and pyDOCK. Most methods also include an atomic repulsion term to avoid steric clashes, either via implementation of molecular mechanics Lennard-Jones potentials (e.g. HADDOCK, RosettaDock and ATTRACT) or by identifying clashes by measuring the molecular overlap (e.g. FFT-based methods and PatchDock[25]). The tendency of molecules to bury part of their solvent accessible area upon interaction is also valuable in discriminating near-native models and is implemented in several docking methods either by valuing large complementary patches (FFT-based and geometric hashing methods) or explicitly via calculation of the buried surface area (e.g. HADDOCK). Related to the buried surface area, several algorithms also incorporate a desolvation energy term that attempts to model the thermodynamic cost or gain associated with changes in the solvation layer of the individual molecules upon binding (e.g. FastContact[40], pyDock, HADDOCK, ZDOCK, and PatchDock/FireDock[41]).

Another common strategy in scoring is to use knowledge-based functions derived from statistical analysis of experimentally determined structures. A recent study[42] identified several knowledge-based scoring functions that seem to discriminate near-native models particularly well: SIPPER[43], DECK[44], and PISA/PIE[45]. Apart from these recently developed functions, others such as ZRANK[46] and DFIRE[47] have been in use for years in the docking community. Interestingly, coarse-grained scoring functions perform remarkably well, in particular for complexes that undergo some conformational changes upon binding (e.g. PIE). Another interesting and successful knowledge-based scoring function is SPIDER[48], developed from an analysis of geometrical networks of residue contacts at protein-protein interfaces. Contrasting with the previous approaches, some research groups have recently developed methods to score protein complexes based on homology and evolutionary information that have performed remarkably well[49], although the idea is not novel[50]. Finally, some methods have also used the entire pool of models generated by the sampling stage to perform a statistical analysis of the pairwise residue contacts and rank the models based on these observations[51]. In general, the application and implementation of these functions is not trivial since they are often tied to a particular docking software or approach. Several scoring functions are available as user-friendly web-servers (e.g. ZRANK and DFIRE) or through a standalone application users can download and execute locally (e.g. DECK, DCOMPLEX[47], PIE/PISA, and SPIDER). Also, the combination of scores from different functions is not necessarily additive, given the redundancy of their individual terms, and requires special attention and expert knowledge[42].

Despite all the development in scoring functions, it is still unwise to choose blindly the best scoring model as the representative of the native state. The assumption that near-native models populate wide low-energy wells in the energy landscape has led to the development and application of clustering algorithms to molecular simulations[52] and, naturally, to docking predictions[53]. ClusPro was the first automated docking method using clustering for near-native model selection[22,54]. The results of the rigid-body sampling stage are clustered using a structural similarity measure that assesses the differences at the interface of the smaller (ligand) molecule (ligand RMSD, used in CAPRI as an evaluation measure[9]) and then the centroids of the clusters are ranked and provided as final models. Other docking

approaches have since implemented clustering algorithms[25,30,55], most using analogous similarity measures, while others have developed similarity measures specific for docking (e.g. fraction of common contacts[56]).

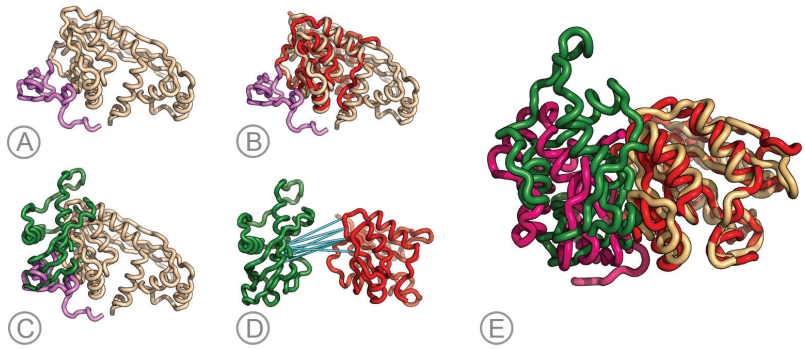


Figure 2. Using information from structural homologues as restraints: HADDOCKing the Mtq2/Trm112 complex. (A) The zinc-binding domain (pink) and methyltransferase (wheat) domain of RlmA(I) (PDB: 1P91 [112]). (B) and (C) Superimposition of the starting models of Mtq2 (red) and Trm112 (green) on the respective distant structural homologous domains of RlmA(I). A simple superimposition of the proteins onto the template would not have produced acceptable models under CAPRI criteria. (D) Distance restraints derived from Cα-Cα contacts between the domains of RlmA(I) mapped onto the Mtq2 and Trm112 models. Only a subset of the 475 unambiguous distance restraints is shown, for simplicity. For a detailed description of the docking parameters refer to *Rodrigues et al., Proteins, 2013*. (E) Submitted HADDOCK model (red/green) superimposed on the native structure (PDB: 3Q87 [108]) (gold/magenta). The interface RMSD. from the native structure is 3.43Å, as calculated by the CAPRI assessors. The restraints, together with the flexible refinement, were instrumental in bringing the model under the 4Å acceptable RMSD. cutoff for CAPRI.

Using experimental information in docking

In an effort to address the limitations of both sampling and scoring, the docking community started developing algorithms that use, somehow, experimental, or predicted information about the interaction (Figure 1) [14,57]. The information is used to bias the sampling and/or scoring. This consequently increases the accuracy of the final models, but depends of course on the quality of the information used. Fortunately, incomplete or partially incorrect information can still be used to great advantage of the prediction[15]. Depending on the use of one or several types of information in the calculations, these methods are designated "data- or information-driven" or "integrative" respectively. How the experimental data are included in the docking calculations differs from method to method, and sometimes between the different stages within a method as well. The simplest implementation is by means of a filter: after generating a number of models, each model is evaluated on its agreement with the provided experimental data, and those models that do not agree or do so poorly are removed or ignored for later stages of the protocol. Another way of incorporating experimental data is through a restraint energy term explicitly included in the sampling. This has the benefit of narrowing the conformational space search and avoids producing models that are in obvious disagreement with the data[58]. Nevertheless, if the data are wrong or irrelevant for the interaction mode under study, the system will explore a wrong region of the interaction space

and native-like solutions will unlikely be sampled. When combined with flexible protocols, direct inclusion of data as restraints might also facilitate the modeling of conformational changes occurring upon binding, albeit with limited efficiency (see Figure 2 in de Vries *et al.*[59]).

The feasibility to implement a certain type of data as a restraint depends, however, on the sampling method. Monte Carlo methods evaluate, at each step, an energy function and use its result to calculate the probability of that particular step being accepted. On the other hand, derivative-based methods such as conjugate gradient energy minimization or molecular-dynamics-based simulated annealing require an energy term that can be described by a continuous function, fully derivable. For example, the buried surface area, calculated empirically using grids, cannot be represented by a differentiable function. Consequently, such functions cannot be implemented in an energy minimization protocol as a restraint, but are perfectly suitable for filter-based approaches or restraints-based approach using non-derivative sampling methods such as Monte Carlo searches or genetic algorithms.

State-of-the-art of integrative modeling software

The benefits of integrating experimental or prediction data into docking are clear in the latest CAPRI assessment report[13]. Supplementary Table 1 summarizes the docking approaches of the five most successful research groups in the "prediction" category of the most recent CAPRI round (HADDOCK, SwarmDock, ClusPro, GRAMM-X[60], and PyDock), as well as other consistently good performers (ZDOCK, ATTRACT, RosettaDock, HEX, PatchDock). This list is by no means complete and the methods described therein are not necessarily the best for every possible case. For a better compilation of the existing docking approaches, refer to the latest CAPRI assessment report[13] and to the review by Moreira *et al.*[57].

All the listed methods offer the possibility to integrate data in their algorithms, although they differ in the stage at which the data can be incorporated and in their implementation. FFT-based approaches (ClusPro, PyDock, GRAMM-X, ZDOCK, HEX) evaluate very large numbers of interaction modes and implement the data as a scoring bias during the FFT search or as filter at the end. PyDock uses the data in a final analysis step, adding an "agreement" score to the models that depends on the type of data. In the case of small angle x-ray scattering (SAXS) data, PyDock uses CRYSOL[61] to generate synthetic scattering curves and then calculate the fit to the experimentally determined curve[62]; in case of distance restraints, PyDock calculates the percentage of user-provided distances respected by the model[63]. ClusPro translates interface data into regions that its search algorithm can either favor or avoid. For the "attractive" restraints, the scoring of the FFT method is changed and the "favored" atoms contribute more to the score if in the interface. Note that this does not exclusively restrict the search to the predefined area; if another surface patch is very energetically favorable (i.e. electrostatic complementarity or hydrophobic packing) then some models might overlook the provided information and prefer this binding mode. A particular feature of ClusPro is the possibility to define "repulsive" regions. These allow the specification of regions of the interacting proteins that should *not* be included in the interface, creating effectively a "negative" restraint that, unlike the "attractive" restraints *must* be respected. Both options are available in the public web server. GRAMM-X also allows the definition of regions of the partners that should be in the interface of the predicted models. The user can also define how many pairs of restraints must be satisfied, thus providing control over the strictness of the filtering algorithm.

Approaches like SwarmDock or HADDOCK use different sampling algorithms and implement the data differently. SwarmDock places the starting molecules with the interfacial residues "in line of sight" of each other, effectively pre-orienting them according to the data. The standard optimization and refinement then take place, without any contribution of the data.

HADDOCK is the pioneer of "data-driven" docking. It philosophically descends from NMR structure calculations[64] and incorporates the data directly into the calculations as an additional restraint energy term. More importantly, the data are used throughout the entire protocol, even during the flexible refinement stages. In HADDOCK, the molecules are first separated in space, their orientation randomized, and then the energy of the system is minimized using a function that includes electrostatics, van der Waals, and the data as distance (e.g. nuclear Overhauser effect, chemical cross-links, electron paramagnetic resonance distances, mutagenesis) and/or orientation (e.g. residual dipolar couplings[65], pseudo-contact shifts[66]) restraints. Other types of data that cannot be directly added to the energy function to be minimized, such as SAXS profiles, are implemented as an additional term in the scoring function[67].

To deal with the high ambiguity of some types of experimental information such as NMR chemical shift perturbations, which defines patches of interacting residues but not the pairwise relationships between them, HADDOCK implements and extends the ambiguous distance restraint concept used in NMR structure calculations[68] as ambiguous interaction restraints (AIRs). These AIRs are essentially one-to-many relationships between residues of different interacting molecules (HADDOCK supports simultaneous docking of up to six molecules[69]) and have the benefit of allowing the sampling method to choose energetically favored binding modes (electrostatics and van der Waals) that agree with (only a portion of) the information.

What truly sets HADDOCK apart from all other docking approaches is the use of the same (or different) data as restraints during the more accurate refinement stages (see Figure 2). In the default HADDOCK protocol, the rigid-body energy minimization stage produces and scores 1000 models (effectively sampling 10000 models – an internal scoring being already performed). The best 200 are then optimized through a semi-flexible simulated annealing in torsion angle space that allows for small conformational changes of up to 2Å[59] (although improvements of up to ~5Å have been observed for flexible peptides[70]), mimicking the induced fit mechanism of binding. The user-provided interface data are also used in this stage, "pulling" the interacting residues towards each other during the flexible optimization of the interface. The final refinement stage – a short restrained molecular dynamic simulation in explicit solvent – also uses the same information. All in all, the use of experimental or predicted interface data in HADDOCK pervades the entire protocol, not only during the orientation of the binding partners as most other data-friendly approaches, but also in the flexible refinement of the modeled conformations.

Besides the top performing approaches in the latest round of CAPRI, other traditionally well-performing approaches are worth mentioning. The coarse-grained ATTRACT docking developed by Zacharias was recently extended to allow docking using cryo-electron microscopy (cryo-EM) density maps[71]. The very efficient docking suite by Wolfson[25] is split in different applications that perform specific tasks: PatchDock generates rigid-body models based on geometric hashing techniques, FlexDock introduces flexibility around hinges, and FireDock refines the docked models. Other specific tools deal with symmetrical assemblies[72,73]. PatchDock can use interface information in a similar way to ClusPro, allowing the

definition of "attractive" and "repulsive" regions, as well as offering the possibility of setting distance constraints. The same group has also developed methods to position proteins in cryo-EM density maps (MULTIFIT)[74]. The RosettaDock approach of Gray makes use of the well-known Rosetta framework and can include experimental data in distance-based filters to bias the Monte Carlo sampling. Users of RosettaDock can also manually position the binding partners in space and then restrict the protocol to sample and refine only around these. The spherical polar Fourier approach HEX by Ritchie has been programmed to make use of graphical processing units (GPUs) and therefore boasts a spectacular speed (~15 seconds for a typical docking run)[75]. It can also make use of experimental data, by specifying residues to define an intermolecular axis around which the rotational search will take place. Finally, ZDOCK, developed by Weng and one of the oldest participants in CAPRI, has been continuously developed and improved over the years, and includes docking-specific knowledge-based scoring functions (e.g. ZRANK, IRAD and IFACE[29]) to complement a exhaustive FFT-based search powered by a new 3D convolution library[76].

Sources of experimental data used in docking

As elaborated before, the most successful docking approaches in CAPRI, as well as many others, recognized the power of adding experimental data into the algorithm to bias the sampling and/or scoring. For experimentalists, this is a tremendous advantage, but one that requires careful planning since bad quality data will produce bad quality models ("*garbage in, garbage out*"). Having data from several sources is useful, but only when the information is not redundant.

Unambiguous high-resolution data on the interface is obviously the best information to use with docking. It maps one atom/residue on a monomer to another atom/residue on another monomer with, or without, a specific distance. Short (up to ~6Å) inter-proton distances from NMR, based on the Nuclear Overhauser Effect, are a classical example. Other experiments are able to provide unambiguous distances over larger distances, such as Förster resonance energy transfer[77], chemical cross-linking[78], electron paramagnetic resonance spectroscopy[79], but these are in many cases of limited accuracy. It is important to note that, although usually distances are used to bring two atoms/residues in contact, they can also be used to keep them apart: if two residues are observed to have a very large minimum distance between each other, this can be valuable information for the docking as well.

Information about the involved interface residues is somewhat easier to obtain and especially powerful when multiple sources are combined. For example, mutagenesis, NMR chemical shift perturbation and cross-saturation, hydrogen/deuterium exchange and limited proteolysis all allow to narrow the interacting surface of the monomers to (small) regions. Most data-driven docking approaches offer support for this sort of ambiguous information, implementing it such that an energetic "bonus" is given to the models that place the selected regions at the interface. This leniency allows a portion of the data not to be respected, which is useful in case of false positives in the experimental data. In particular, HADDOCK randomly removes half (by default – can be defined) of the ambiguous interaction restraints for each docking trial. This means that each model is generated with a slightly different set of restraints, decreasing the odds that erroneous information forces the docking into sampling and scoring wrong conformations.

Shape information from SAXS, SANS (small angle neutron scattering), and cryo-EM also provide valuable data, in particular for asymmetric complexes and assemblies of many

partners. Although not trivial to implement, explicitly using density maps obtained from cryo-EM and fitting the interacting molecules into them is extremely helpful to reduce the conformational search. This of course depends on the resolution of the density maps. The same is valid for SAXS and SANS. The resulting scattering curves, or the radius of gyration derived from them, can be used for scoring. The latter can also directly be used as a restraint in docking. SANS, in particular, has been shown to be specifically useful for protein-nucleic acid complexes, as shown in a recent study on a large molecular weight protein-protein-RNA complex[80]. Collision cross section (CSS) data obtained from ion mobility mass spectrometry (IM-MS)[81] can also be valuable for describing protein complex shape, although the information content for smaller systems might be rather limited[62,67].

Predicting from predictions: bioinformatics interface predictions

Biological systems such as protein-protein complexes are not always researcher-friendly and more often than not refuse to behave under the experimental conditions required to obtain structural data. In such cases, bioinformatics prediction methods can provide hints that can be used instead of experimental information to drive the docking calculations (Supplementary Table 2). Predictions are also extremely useful when only a tiny amount of experimental information is available. Much like experimental techniques, bioinformatics interface predictors provide different types of data depending on their algorithms and theoretical principles. Most methods use sequence/structure conservation to define residues that might be important for interactions (e.g. WHISCY[82], PredUs[83], Consurf[84]). Other methods analyze structural properties of the monomers, such as surface exposed residues, and/or employ propensity-scales for pairs of residues in protein complexes (e.g. InterProSurf[85], SPPIDER[86], ProMate[87]). Others yet, combine both types of information or several different methods to offer a consensus or meta-prediction, which has been shown to be more accurate than the individual methods (e.g. CPORT[88], PresCont[89], meta-PPISP[90]).

Surprisingly, most prediction servers take into account only the monomer sequence and/or structure for their analysis: i.e. the predictions are not partner-specific. The predictions tend then to be hardly specific and a give rise to large and diffuse potential interaction patches on the surface of the monomers. To increase specificity, some research groups developed partner-aware or partner-specific methods[91,92]. Motivated by the colossal amount of sequence information from mass sequencing efforts, methods have also been developed to dig deep in the evolutionary history of protein sequences hidden in multiple sequence alignments. By correlating residues across potentially interacting sequences, unambiguous interaction information at the residue level can be obtained. This has been used, so far, to score the models[49,50], but recent developments hint at a potential application in other stages of docking[93,94].

Last but not least, docking programs have been used to predict binding regions for docking programs. The idea that given enough sampling and an energy function, the true interface will be sampled more often than at random has garnered quite a following, despite the obvious deficiencies of sampling/scoring mentioned at the beginning of this manuscript. Nevertheless, some of these approaches (e.g. RCF/ZDOCK[95]) have been applied in the latest CAPRI round and shown interesting results that hint that statistical analysis of contacts in large pools of models have some degree of predictive information. Still, this only applies for very large pools of models where the sampling is exhaustive, and relies exclusively on energy functions that have to be simple to assess such number of models in a reasonable

period of time. As a rule of thumb, and as some of these prediction servers openly advise, the predictions should be analyzed carefully and critically, preferably in light of some experimental or functional data.

Beyond docking: other approaches for protein interaction modeling

Docking is not the only method to model protein interactions since, in principle, any structure prediction can be applied to the docking problem. Among these, a few particular methods deserve praise: the Integrative Modeling Platform (IMP)[96] and the modeling of complexes based on homologues[97-99].

IMP is a multi-scale umbrella approach that combines several different scripts to model structures from various sources of information, most notably cryo-EM and SAXS. It is likely the current method of choice for modeling very large assemblies, since most docking methods are limited to a few interacting partners and very few support the low-resolution data that helps narrowing the conformational space search in such cases.

Another class of approaches to model protein complexes uses existing structures of homologous complexes to define the binding mode. Although modeling individual monomers by homology has been widely used as a step prior to docking[15], there is also a wealth of information in experimentally determined structures of homologous protein complexes. PRISM was developed to search and find, by rigid-body structural comparisons, template protein-protein interfaces that match a target protein, and then refine it using flexible docking protocols[97]. COTH[98] is a powerful iterative threading approach that has been developed for protein-protein complex prediction: it first queries a database of complex sequences and finds the ten best matches, proceeds to model the sequence of the individual monomers, and finally superimposes these on the ten original complex matches. Both methods were benchmarked against ZDOCK and a novel homology-based method (ZTEM) in a recent publication[100]. Kundrotas *et al.* claimed that "*there are enough templates to model nearly all protein complexes*"[99] and developed a method to rigidly superimpose the interacting proteins to homologous complexes. Finally, KBDOCK was also recently developed and uses domain-domain homology to model protein complexes[101].

Yet, sometimes even close homologues interact differently[102] or show completely different specificities[103]. Homology-based methods will never accurately describe such effects, nor allow the modeling of novel interaction modes. Additionally, the template libraries used in these approaches should be expanded to include multi-domain proteins, often the evolutionary forefathers of protein complexes, as demonstrated by the (only) successful prediction of the complex between the methyltransferase MTq2 and an activator protein (Trm12) from *E. cuniculi* (PDB: 3Q87[104]) by the HADDOCK group[15].

Concluding remarks

Protein docking emerged in the late seventies after a decade of methodological leaps in computational biology[7,105]. It has since grown, come of age, and established itself as a pivotal method in structural biology. The accuracy of current docking methods is far from perfect, but the inclusion of experimental information has benefited them immensely. These so-called data-driven or integrative modeling methods can nowadays provide useful hints or even definite answers to biological problems. Their ability to deal with several sources of data and tackle evermore-complex systems is growing steadily. Although, plenty of challenges remain,

e.g. genome-wide prediction, large conformational changes, binding-affinity prediction, there is progress in all fronts with quite satisfactory results[106]. While it is unreasonable to state that perfect predictions are around the corner, it is perfectly realistic to wonder if the next generation of scientists will be, much like their methods, computational/experimental hybrids and if these hold the key to unravel the mysteries of cellular and molecular biology.

Supplementary Tables

Supplementary Table 1. List of top-performing docking approaches participating in CAPRI

Name	Protocol	Strengths & Weaknesses	Integration of Data	Public Web Server
ATTRACT	Energy minimization in translational and rotational degrees of freedom using a reduced protein model and normal-mode analysis to allow conformational changes upon binding.	+ Fast derivative-based search method + Conformational changes upon binding (local and global motions) + Support for Cryo-EM density maps - Not available via a web server	Implements interface data as by adding atom/residue specific weights, which can be negative (repulsive). Also offers the option to dock using Cryo-EM density maps.	None
ClusPro	Rigid-body search via a FFT correlation approach (PIPER), followed by structural similarity (RMSD) based clustering to find the most popular interaction modes, and final refinement of selected structures using CHARMM.	+ Best automated served in the latest CAPRI evaluation [12]. + Fast and exhaustive protocol + Several docking 'modes' depending on the biological function (antibody/antigen, multimer, others). - Cannot handle flexible complexes.	Regions/Residues in the binding partners can be introduced to bias the scoring of the models. Noteworthy option of 'negative' (repulsion) contacts.	http://cluspro.bu.edu/
GRAMM-X*	Grid-based FFT rigid body docking approach using a softened Lennard-Jones potential function. The top predictions are minimized and re-scored using a soft Lennard-Jones potential, an evolutionary conservation term, a statistical residue-residue potential, and the volume of the local energy minima in the grid.	+ Fast + Uses CONSURF to determine evolutionary conserved residues. - Cannot handle flexible complexes.	Regions/Residues in the binding partners can be introduced to bias the scoring of the models.	http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/
HADDOCK	Rigid-body energy minimization followed by semi-flexible (interface) refinement and final optimization in explicit solvent. Returns clusters of models ranked by HADDOCK score.	+ Best performing team in the latest CAPRI evaluation [12]. + Restraint-based integration of ambiguous experimental/prediction data. + Explicit flexibility of the interface. + Powerful user-friendly web interface. - Slower than FFT-based methods.	Several types of restraints allow integration of different sources of data: distances, orientations, radius of gyration, symmetry type, etc. Directly integrates several NMR data.	http://haddocking.org
Hex	Spherical polar Fourier approach using rotational correlations generates a very large number of putative models, which are then re-scored using a shape correlations or shape plus electrostatic correlations.	+ Extremely fast approach (~15s using 2 GPUs). - Cannot handle flexible complexes.	Can restrict the rotational search around an intermolecular axis defined by a pair of residues, one on each interacting monomer. The angular range of the search can also be defined.	http://hexserver.loria.fr/

* The research group responsible for GRAMM-X has recently developed an alternative technique to predict the structure of protein-protein complexes based on structural similarity with (distant) homologues. The method is reviewed in greater detail in the 'Beyond Docking' section of this manuscript.

Supplementary Table 1 (continued). List of top-performing docking approaches participating in CAPRI

Name	Protocol	Strengths & Weaknesses	Integration of Data	Public Web Server
PatchDock	The surface of the molecules is divided in patches (concave, convex, and flat) and only those containing 'hot-spot' residues are kept. The patches are then matched using geometric hashing and pose-clustering techniques and the candidate models are examined (to remove extreme clashes) and scored.	+ Extremely efficient and fast protocol. + Integrated suite of docking tools [24]. - Fragmentation of protocol in very specialized tools requires a priori knowledge of their limitations.	Regions/Residues in the binding partners can be introduced to bias the scoring of the models.	http://bioinfo3d.cs.tau.ac.il/PatchDock/
PyDock	Rigid-body search via a FFT search method (custom optimized FTDOCK) followed by scoring with a combined electrostatics and desolvation energy function.	+ Fast protocol. + Integration of different modules (pyDockSIPPER, pyDockRST, pyDockSAXS) to improve predictions. - Cannot handle flexible complexes.	PyDockRST module scores models based on agreement with user-defined distances. Implements SAXS scoring by generating synthetic SAXS curves of the models to user-provided data.	http://life.bsc.es/servlet/pydock/home/
RosettaDock	Low-resolution, rigid-body, Monte Carlo search followed by simultaneous optimization of backbone displacement and side-chain conformations using Monte Carlo minimization.	+ Scoring based on the Rosetta energy function + Very powerful refinement protocol - Poor (direct) support of interface information - Computationally demanding and slow protocol.	Molecules can be positioned manually in space (e.g. using PyMOL). Interfaces with other Rosetta tools (RosettaInterface) to validate mutagenesis data.	http://antibody.graylab.jhu.edu/docking
SwarmDock	Local docking and particle swarm optimization of partner position and orientation, using normal modes to model induced fit, and final energy minimization. Uses the DComplex scoring function but the final models are re-ranked with a centroid potential prior to clustering.	+ Search method that explicitly models global flexibility upon binding. - Slow. Predictions can take days unless a local version and sufficient computational resources are available.	Residues belonging to the binding site can be selected to bias the starting positions of the binding partners.	http://bmm.cancerresearchuk.org/~SwarmDock
ZDOCK	FFT-based rigid-body search using a scoring function composed of desolvation energy, electrostatics, and grid-based shape complementarity. Recent versions implement knowledge-based scoring functions and clustering to analyze the energy landscape of binding.	+ Fast protocol + Robust performance over several CAPRI rounds. - Cannot handle flexible complexes.	Allows user-defined selection of 'blocked' and 'binding' residues that influence the scoring of the FFT-based search.	http://zdock.umassmed.edu/

Supplementary Table 2. List of interface prediction methods available for docking predictions*

Name	Type of Prediction	Public Web Server
ConSurf	Identifies close sequence homologues using (PSI-)BLAST, builds a multiple sequence alignment and then a phylogenetic tree. Position-specific conservation scores are then calculated using a empirical Bayesian or Maximum-Likelihood algorithm and divided into a discrete scale of nine grades.	http://consurf.tau.ac.il/
CPORT	Consensus sequence-/structure-based predictor that combines WHISCY, PIER, ProMate, cons-PPISP, SPPIDER, and PINUP.	http://haddock.chem.uu.nl/services/CPORT
InterProSurf	Uses the solvent accessible area of the monomers, together with a propensity scale for interface residues, and a clustering algorithm to identify high-scoring patches on the protein surface.	http://curie.utmb.edu/prosurf.html
meta-PPISP	Consensus structure-based predictor that combines cons-PPISP, PINUP, and ProMate.	http://pipe.scs.fsu.edu/meta-ppisp.html
PredUs	Potential interface residues are identified by iteratively mapping interaction sites of close and remote structural neighbors to individual residues on the query protein.	http://bhapp.c2b2.columbia.edu/PredUs/
PresCont	Uses four residues properties – solvent accessible area, hydrophobicity, sequence conservation, and local environment of the amino acid in the protein – in a support vector machine classifier.	http://www-bioinf.uni-regensburg.de/php/prescont.php
ProMate	Analyses the chemical character of surface residues, such as clustered hydrophobic and polar amino acids, as well as the B-factor of the residues in the unbound state.	http://bioinfo.weizmann.ac.il/promate/
PS-HomPPI	Partner-specific interface predictor based on the *k* nearest interologues.	http://einstein.cs.iastate.edu/PSHOMPPI/
RCF	Uses ZDOCK to generate models of the interaction and then analyses the contact frequency of each residue for each partner.	None
SPPIDER	Integrates enhanced relative solvent accessibility predictions, evolutionary information, high-resolution structural data, and physico-chemical properties in a machine learning approach.	http://sppider.cchmc.org/
WHISCY	Sequence conservation aided by surface structural information and the propensity scale for interface residues.	http://nmr.chem.uu.nl/Software/whiscy/

* For a complete overview of the available methods for protein-protein interface prediction and their performance refer to more in-depth reviews such as [107] and [108]

Chapter 2

Defining the limits of homology modelling in information-driven protein docking

J. Rodrigues, A. Melquiond, E. Karaca, M. Trellet, M. van Dijk, G. van Zundert, C. Schmitz, S. de Vries, A. Bordogna, L. Bonati, P.Kastritis and A.M.J.J. Bonvin

Published in 2013 in Proteins, volume 81, issue 12

Abstract

Information-driven docking is currently one of the most successful approaches to obtain structural models of protein interactions as demonstrated in the latest round of CAPRI. While various experimental and computational techniques can be used to retrieve information about the binding mode, the availability of three-dimensional structures of the interacting partners remains a limiting factor. Fortunately, the wealth of structural information gathered by large-scale initiatives allows for homology-based modelling of a significant fraction of the protein universe. Defining the limits of information-driven docking based on such homology models is therefore highly relevant. Here we show using previous CAPRI targets, that out of a variety of measures, the global sequence identity between template and target is a simple but reliable predictor of the achievable quality of the docking models. This indicates that a well-defined overall fold is critical for the interaction. Furthermore, the quality of the data at our disposal to characterize the interaction plays a determinant role in the success of the docking. Given reliable interface information we can obtain acceptable predictions even at low global sequence identity. These results, which define the boundaries between trustworthy and unreliable predictions, should guide both experts and non-experts in defining the limits of what is achievable by docking. This is highly relevant considering that the fraction of the interactome amenable for docking is only bound to grow as the number of experimentally solved structures increases.

Introduction

Bruce Alberts referred to the cell as a 'collection of protein machines'[109]. This simple definition masks the complexity behind the machinery that indeed governs all processes essential to life. Much like most machines interface with others of their kind to collaboratively achieve a greater goal, proteins in the cell are organized in pathways, or networks, which are regulated with a formidable complexity[110]. These networks, collectively called the 'interactome', are the fabric of life itself. Unfortunately, and despite decades of research, a large fraction of the interactome remains in the dark, unknown, and therefore beyond our reach[106]. While cellular biology and molecular biology often answer the 'what' and 'where', knowledge of 'how' a specific network operates begs for high-resolution structural information. Yet, experimental structural characterization of protein interactions is progressing slowly compared to our increasing knowledge of the interactome. At the same time, we have access to a wealth of information that could potentially be used in computational structure prediction algorithms[99,106].

The prediction of the structure of a protein-protein complex *in silico* is not novel[105] and nowadays can be carried out using one of two major methods: comparative modelling and computational docking. Comparative modelling relies on the notion that a pair of interologs (conserved interaction between a pair of proteins which have interacting homologs in another organism) often shares the same binding interface[102]. However, this approach can only reliably target a fraction of the interaction space: interactions for which no interologs can be found, or for which those found are below the threshold of reliability, cannot be modeled by comparative methods. Also, sequence similarity does not always convey interaction similarity[111], nor even interaction specificity, as illustrated by a recent study on enzymes of the ubiquitination pathway[103]. In contrast to comparative modelling, computational docking predicts the structure of protein interactions from the structures of the unbound interacting partners by performing a search in the interaction space and assessing each model based on some scoring function. Community efforts on blind predictions (CAPRI[9-11,112]) have shown that explicit integration of experimental information during the docking calculations is valuable and increases their accuracy considerably[14,112].

Regardless of the approach chosen, there is always the need for a three-dimensional (3D) structure of the interacting partners to start the modelling process. Large-scale structural genomics initiatives such as the Protein Structure Initiative[113] make it possible, to a certain degree, to find a sequence homolog with known structure, which can then be used to build a 3D model of the protein of interest. This, combined with the availability of homology modelling algorithms through web interfaces[114,115], makes it rather simple for non-experts to build models that can serve as input for docking predictions. However, simply put, a completely wrong model will never yield a good prediction. Akin to the notion of 'twilight zone' of sequence alignment for homology modelling[116], which defines the sequence identity/similarity limit from which one can expect to build a reliable model, there must be an equally important 'zone' where homology models are suitable for docking. The definition of this 'twilight zone' for protein interaction modelling from homology models is therefore critical for single docking predictions and, perhaps more importantly, for high-throughput predictions of entire interactomes.

In this work, we address these concerns and identify the most suitable predictive metric for the reliability of homology-based information-driven docking. Using previous CAPRI targets for which information-driven docking has proven successful[10,11,112], we generated by

homology modelling structural models of varying sequence identities and perform docking with HADDOCK[30,117,118], using the same information used in CAPRI. We analyze several sequence- and structure-based metrics, and discuss the impact of the quality of the homology model on the information-driven docking prediction. The influence of the quality of the interaction data on the final models is also analyzed. This allows us to define the limits of the achievable quality of a docking model for a given quality of a homology model. These are independently corroborated by our prediction results obtained in the last CAPRI rounds, which are also shortly discussed and summarized here.

Materials and Methods

Dataset of Protein-Protein complexes

To assess the impact of the quality of a homology model in information-driven docking, we used protein-protein complexes from previous CAPRI targets (from rounds 4 to 19, Target 12 to Target 42) for which we had obtained a successful prediction. We only considered complexes for which at least one of the interacting partners has sequence homologs with an experimentally determined structure. These represent "real-life" scenarios. Protein-protein complexes presented in the last rounds of CAPRI were used as an independent validation set. The complexes that met the criteria and those used for validation are listed in Table 1.

Homology modelling of interacting partners

Homology modelling was performed using a simple and straightforward protocol, detailed in the Supplementary Material. Thirty models per interacting partner/template pair (10 per alignment method) were generated, resulting in a total of 870 different models across the entire dataset (for a total of 29 homologues for 6 targets – see Table 1). Unaligned regions that resulted in long disordered loops or termini were removed.

Information-driven docking predictions using HADDOCK

We performed docking predictions for all targets using the various homology models of each chain and the reference bound structure of the other partner. In addition, for each target, bound-bound docking was performed to measure the best possible outcome. Homology model – homology model docking was not performed since it would be more difficult to isolate the impact of a given model on the docking results quality. This scenario is however present in the independent set from the current CAPRI rounds.

Two sets of restraints were used for each run: *CAPRI interface restraints (CI) and true interface restraints (TI)*. The CAPRI restraints comprise the information used during the corresponding CAPRI round, which is described in detail in previous publications[117,127] and is available upon request. True interface restraints were derived from the reference structure of the complex (all residues on each chain at a minimal atom distance cut-off of 5Å from any residue in the other interacting partner), and included as ambiguous interaction restraints (AIRs).

All docking predictions were performed with HADDOCK[30,117] (beta version 2.2), using CNS[128] (version 1.3) for the structure calculations, with default settings, except for the number of models which was set to match our previous CAPRI submissions. The HADDOCK score (details in the Supplementary Information), used to rank the generated models after water refinement, consists of a weighted sum of physics-based energy terms (electrostatics and van der Waals) complemented by an empirical desolvation energy term (E_{desolv})[129] and

a restraints energy term (E_{AIR}), as defined in the following equation:

$$\text{HADDOCK Score} = 0.2 \times E_{elec} + 1.0 \times E_{vdW} + 1.0 \times E_{desolv} + 0.1 \times E_{AIR}$$

Assessment of structure similarity and docking predictions

To compare the structures of the homology models to those of the native proteins, individually or in the complex, two metrics based on the root mean square deviation of atomic coordinates (RMSD) were used:

backbone RMSD (bbRMSD): calculated between two chains and on the backbone atoms of the molecules (Cα, C, N, O).

interface RMSD (i-RMSD): as defined by CAPRI[9], calculated on the backbone atoms of residues within a minimal atom distance cut-off of 10Å of any residues of a different molecule of the complex. When comparing single homology models to the crystal structure, this metric refers only to the interfacial backbone atoms of that partner.

To assess the quality of the restraints we calculated both precision and recall as defined in the equations below:

$$\text{Precision} = \frac{\#\text{ of correctly predicted interface residues}}{\#\text{ of predicted interface residues}}$$

$$\text{Recall} = \frac{\#\text{ of correctly predicted interface residues}}{\#\text{ of interface residues in crystal}}$$

Predictive and measured indices for model quality

A previous study on the impact of homology models in protein-ligand docking[130] used several sequence- and structure-based indices to assess the quality of the models on the docking predictions. These are divided into two categories: 'predictive' and 'measured' indices. Predictive indices are those that can be calculated without prior knowledge of the structure of the complex, and can, therefore, be used to estimate the success of the docking prediction from the homology models. Measured indices are calculated knowing the structure of the complex and are used here to define the quality of a docking prediction. See Supplementary Material Table S1 for details.

Table 1. Dataset collected for measuring the impact of homology models on the docking predictions.

CAPRI Target Number	PDB ID	Protein Name	No. of Homologues	Sequence Identity of Homologues (%)
Analysis Set				
T12[119]	1OHZ	Cohesin	3	31-71
		Dockerin	2	37-46
T18[120]	1T6G	Xylanase	0	--
		TAXI	4	37-51
T26[121]	2HQS	TolB	0	--
		Pal	4	21-69
T27[122]	2O25	E2-25K	3	22-39
		Ubc9	4	29-55
T40[123]	3E8L	Serine proteinase inhibitor A	3	34-82
		Cationic Trypsin	0	--
T41[124]	2WPT	Im2 immunity protein	4	50-63
		Colicin-E9	2	66-67
Validation Set				
T46[104]	3Q87	Trm112p-like protein	1	19
		Methyltransferase small domain	1	12
T50[125]	3R2X	Influenza Hemagglutinin	0	--
		HB36.3 designed protein	1	85
T53[126]	n/a	Rep4	1	67
		Rep2	0	--

Results

Correlation of predictive indices with docking model quality

We first performed a correlation analysis in order to assess if a particular index showed a predictive trend with respect to the quality of the final docking model (Figure 1A, Supplementary Tables S1 & S2). All indices are described in detail in the Supplementary Material. Sequence-based indices (e.g. sequence identity over the entire sequence or on the interface only) show relatively high and uniform correlations (~0.7 absolute Spearman rank correlation coefficient). Structure-based indices, on the other hand, show a larger degree of heterogeneity, with coefficients ranging from ~0.3 to ~0.7. Overall, the highest coefficients are observed for the TVSMod_RMSD (0.73)[131], backbone i-RMSD between model and template (0.73), and global sequence identity (0.70). The Qmean[132] and Verify3D[133] indices show the lowest correlation coefficients of all (0.39 and 0.37 respectively). The remaining indices have coefficients that fluctuate around 0.60.

We analyzed in more detail a representative of the sequence-based indices and another based on the structural properties of the model. For the sequence-based indices, we opted for sequence identity as it is easily derived from the alignment to be used in the modelling.

As expected, as the sequence identity decreases, the quality of the model worsens (Figure 1B). Interestingly, even at very low identities, well inside the 'twilight zone' of traditional homology modelling (~30% identity), our information-driven docking approach still produces near-native models (<3.5Å i-RMSD). The best correlating of the structure-based indices is TVSMod_RMSD[131]. It is based on support vector machine regression models and aims at predicting the RMSD and the fraction of Cα atoms of the model within 3.5Å of those of the native structure after rigid superimposition. TVSMod_RMSD shows a similar trend as the sequence identity, although seemingly more discriminatory at lower i-RMSD values (Figure 1C): Models that are predicted to be within 2Å of the native structure produce, in general, docked models within 2.5Å of the native complex. Beyond the 2Å prediction, the correlation is less well defined, although there is still an observable trend.

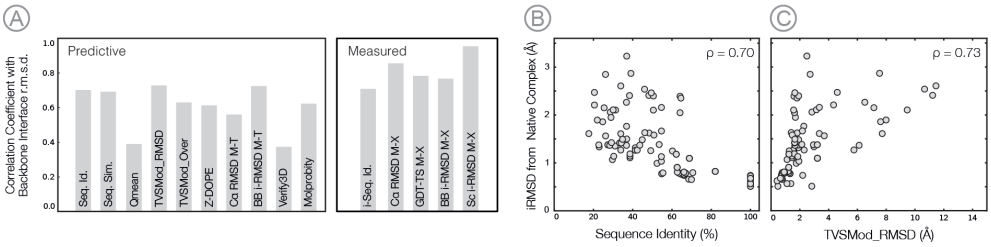


Figure 1. Correlation of sequence- and structure-based indices of homology models with docking model quality (see Supplementary Material for a description and references of the indices used). **A)** Absolute Spearman correlation coefficients of the various indices with the backbone i-RMSD from the native complex structure. **B)** Correlation plot of sequence identity between target and template with the i-RMSD of the best model produced using true interface restraints. **C)** Correlation plot of TVSMod_RMSD score of the homology model and the i-RMSD of the best model produced using true interface restraints. The corresponding Spearman rank correlation coefficients (ρ) are indicated in the figures.

Assessment of the impact of the quality of the data in the docking calculation

Although the quality of the homology model plays a large role in defining the success of the docking calculation, in information-driven docking the quality of the data is also quite relevant. To quantify this, we ran docking calculations with both perfect interface definition, derived from the crystal structure, and interface definitions as obtained from literature and/or bioinformatics predictions during the CAPRI round that produced the target (i.e. a reflection of what a researcher would have at hand in a real-case scenario).

As expected, true interface restraints (meaning that the correct interface residues were used to define ambiguous interaction restraints, which does not restrain the relative orientation of the molecules) produced very accurate results, with all models under 3.5Å i-RMSD of the native complex structure (Figure 2C and 2E). Models produced using CAPRI restraints however, are highly dependent on the quality of the used information (Figure 2B and 2D). We calculated the precision and recall of the collected interface information with respect to the native interface (Figure 2A). The precision of the information across all targets was very high (above 80%), meaning that most of the data used to drive the docking were correct (these had been obtained from literature and bioinformatics predictions[88,117,127]). The recall was also reasonably high (in general above 50%) meaning that the fraction of the

interface that was targeted was sufficient to avoid ambiguity during the calculations. The combination of these factors contributed to a good success rate and low i-RMSD values for most targets. The exception was T18, in which the interface information for chain B was very narrow (recall 0.05), while for chain A the values of both precision and recall – 0.38 and 0.31 respectively – were low. This led to docking solutions that were in general worse than for the other targets. We also assessed if the quality of the homology model, measured by the sequence identity of the template to the target sequence, had a large impact on the ranking of the docking models at the last refinement stage in HADDOCK (Supplementary Figure S1). Again we observed that the quality of the data plays a more important role than the quality of the homology model, since the ranking of the final solutions is largely independent of the sequence identity.

Correlation of measured indices with docking model quality

Besides calculating predictive indices, which can be used to assess *a priori* whether a homology model will be able to produce a good docking model or not, we calculated a series of measured metrics that compare the individual models with the bound structure in the complex (Figure 1A). These, of course, cannot be used without *a priori* knowledge of the native complex. The structure-based metrics were: Cα RMSD, backbone i-RMSD, side-chain i-RMSD, and GDT-TS. All had high correlation coefficients with the quality of the docking models as measured by i-RMSD, with the highest value found for the side-chain i-RMSD of the model to the native complex (0.94). The correlation coefficient obtained for backbone i-RMSD was the lowest of all four (0.69) but still reasonably high when compared with the predictive indices. We also calculated the sequence identity at the interface to assess its importance in defining the quality of the final models. Surprisingly, it shows only a slightly better correlation (0.71) compared to global sequence identity, which can be calculated without knowledge of the native complex.

Impact of the flexible refinement on the quality of the docking predictions

The information used to drive the docking is translated generally into ambiguous distance restraints between residues on the surface of the proteins. This has particular importance during the flexible refinement stage, since interface residues are granted larger freedom. An analysis of the difference in i-RMSD from the native structure between the initial rigid-body docking models and the final refined models reveals modest improvements up to about 1Å RMSD changes (Supplementary Figure S2 A & B). A large fraction of models, however, does not improve or even deteriorates slightly, as is typically observed in molecular dynamics simulations. Interestingly, the best improvements belong to cases with low template identity (Supplementary Figure S2 C & D). The quality of the restraints also plays a role in the extent of the improvement: using true interface restraints shifts the distribution of the differences in i-RMSD between initial and docked model toward more negative (better) values and in general larger improvements. In these cases, the restraints were thus instrumental in improving the model or preventing it from deviating from the correct conformation. In general, these observations are in line with previous studies showing that the impact of flexible refinement is limited with typical maximal improvements in the order of 1.5 to 2.0Å (see Figure 2 in de Vries et al.[117]).

Table 2. Performance of HADDOCK in recent CAPRI rounds. T52 was cancelled and T55 and T56 were special scoring experiments, discussed in detail elsewhere[45].

Target Name	PDB Code	Target Type	Manual Submission Performance[b]	Manual ***/**/*	Server ***/**/*	Uploaded[c] ***/**/*	Scoring ***/**/*
T46[104]	3Q87	HH		0/0/3	0/0/10	0/0/22	0/0/2
T47[134]	3U43	UU		4/6/0/	9/1/0	112/88/0	9/1/0
T48[135]	N/A	UU		0/0/2	0/0/0	0/0/28	--[a]
T49[135]	N/A	UU		0/0/1	0/0/3	0/1/30	--[a]
T50[125]	3R2X	HU		0/2/0	0/0/0	0/15/10	0/0/2
T51[136]	N/A	UUHU(U)		0/2/0	0/0/0	0/1/7	0/0/0
T53[126]	N/A	UH		0/3/1	0/0/0	0/3/43	0/3/5
T54[137]	N/A	UH		0/0/0	0/0/0	0/0/1	0/0/0
T57[138]	4AK2	UU		0/1/3	0/1/1	N/A	N/A
T58[139]	4G9S	UU		0/0/1	0/0/0	N/A	N/A

[a] HADDOCK did not participate in this scoring round.
[b] Quality of the submitted models: high (black), medium (dark grey), acceptable (light grey), incorrect (white).
[c] Uploaded structures: 200 models comprising the best 100 models in both manual and server predictions.
[d] Scoring structures: manual submission of 10 best HADDOCK scored models from scoring set (all groups).

Performance of information-driven docking with homology models in recent CAPRI rounds

The recent CAPRI rounds (22 to 27) provided a wealth of targets that required homology modelling with half of these requiring modelling of at least one of the binding partners. They can thus serve as an independent dataset. Table II reports the success rate of HADDOCK in these last rounds. Out of ten targets, HADDOCK was successful in nine of them, corresponding to an unequalled 90% success rate when considering the manual submission entries. The HADDOCK server automated submission successfully predicted four out of ten targets (40% success rate), but interestingly, in three of these targets (T46, T47, and T49) it outperformed the manual submission. In one, T48, the interaction information provided by the organizers (low-resolution SAXS data) did not improve the scoring of the models. In fact, the lowest fit to the SAXS data, with a χ^2 value close to that of the native crystal structure, belonged to an extremely different and wrong model. Instead, the application of a novel hydrophobicity potential[140] in the standard solvated docking algorithm of HADDOCK[141,142] yielded the only acceptable solution submitted for evaluation.

Overall, these results rank HADDOCK as one of the best docking software to participate in CAPRI (and the best one in this round), reinforcing the idea that data-driven docking is a very successful approach to model biomolecular interactions.

Of the five targets requiring homology modelling, we could only analyze three, as the crystal structures are not yet all publicly available (Table II). The sequence identities of the models produced for CAPRI targets T46, T50, and T53 correlate nicely with the final quality of our best model (Figure 3). The precision and recall rates for the information used to drive the docking were also extremely high (Figure 3), in particular for target T46, in

which we produced the only one-star models by CAPRI criteria (<4Å i-RMSD) despite the very low sequence identity to the templates used for the modelling: 12 and 18 % (refer to the Supplementary Material for a description of the restraints used in this target).

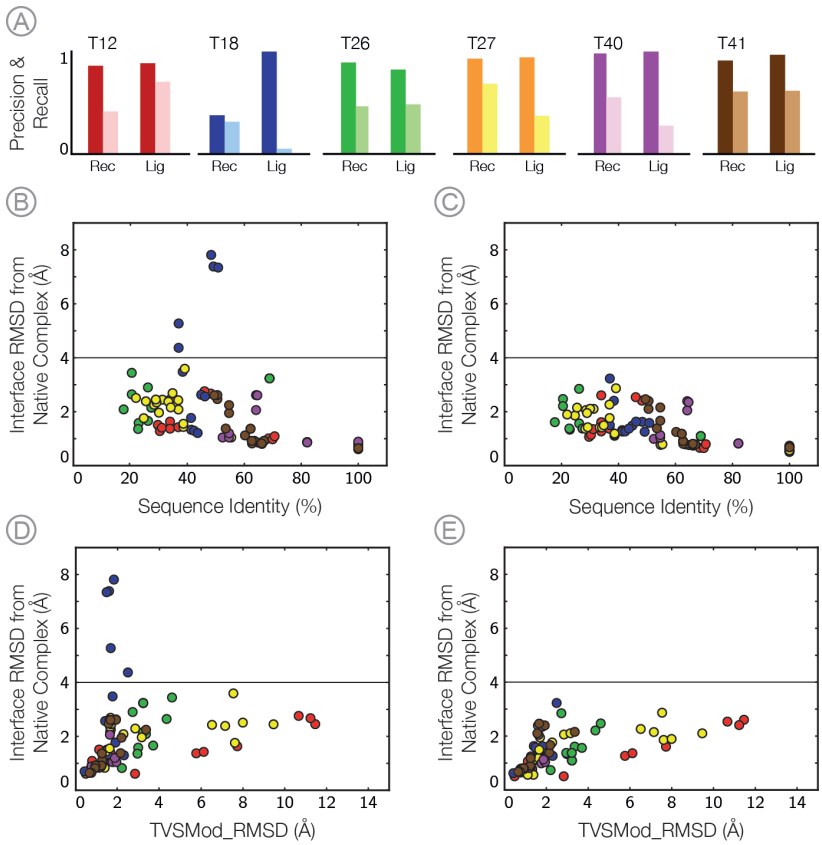


Figure 2. Influence of the quality of the interface data on the docking results. **A)** Precision and recall (see Methods) for the interface information used as restraints (color-coded by CAPRI target). **B & D)** Correlation plot of percentage sequence identity between the model and the template and the TVS-Mod_RMSD score of the homology model with the i RMSD from the native complex of the best docking solution based on CAPRI restraints. **C & E)** Correlation plot of percentage of sequence identity between the model and template and the TVSMod_RMSD score of the homology model and the i RMSD from the native complex of the best docking solution based on true interface restraints

Discussion

Sequence identity is a simple yet reasonably accurate predictor of docking success

The analysis of all sequence- and structure-based indices showed that none performs significantly better than the others. Sequence identity and similarity perform equally well (correlation coefficients of 0.70 and 0.69 respectively) and are trivial to calculate, requiring no further information than the pairwise alignment. Interestingly, the sequence identity at the interface is only a marginally better predictor (correlation coefficient of 0.71), which suggests that the overall fold of the molecule is relevant for a good arrangement of the interface and thus for the success of the docking. Structure-based indices show a rather heterogeneous performance. The QMean[132], Molprobity[143], and Verify3D[133] metrics all evaluate the structural properties of the model, such as amino acid packing, distribution of torsion angles, etc. (Supplementary Table S1). Since the homology models undergo a slight refinement, it is not expected that they have severe clashes or other deviant structural features. Nevertheless, Molprobity was very discriminative of native structures, attributing to these very low scores (almost always below 15 a.u.) in contrast to scores above 70 for the majority of the homology models. The scoring between the homology models was, however, heterogeneous and did not correlate with the docking results. Finally, the backbone i-RMSD between model and template, a direct structural comparison measure, showed the highest correlation coefficient, on par with TVSMod_RMSD (0.73), and better than the overall structural similarity between the two structures (0.56).

The quality of the interaction restraints has a greater impact than the quality of the homology model

Information-driven docking narrows the conformational landscape of association of the molecules to the fraction that respects that information. Furthermore, if the information is integrated in the energy function used in refinement (i.e. not only for scoring), there is an added benefit of driving the interface refinement. Our results are in agreement with these assumptions, since docking calculations using literature-based information (CAPRI restraints, Figure 2B and 2D) show worse results than those using true interface restraints (Figure 2C and 2E). The impact of the quality of the restraints is illustrated in the runs of T18, where precision and recall values were extremely low and the models were accordingly of bad quality (i-RMSD over 4Å). Overall however, despite starting the modelling process with templates as low as 20% sequence identity, the docked models are still quite reasonable (within 3Å i-RMSD), provided that the interaction information is reliable. This thus stresses the importance of the quality of the data over that of the model. The scoring of the models, helped by the interface information, is also robust enough to discriminate good quality models, regardless of the identity of the template used in the homology modelling. This again reinforces the notion that the quality of the data is more important than that of the model, since good data can refine a bad model and discriminate which solutions are closer to the native structure, while weak data pollute the docking protocol even when the model quality is reasonable.

Defining the limits of homology modelling in information-driven docking

Based on these observations, we can predict the quality of information-driven docking predictions given the sequence identity of the templates used to build the homology models (Figure 3). Assuming reliable interface information, a homology model built with a template

sharing 20% sequence identity can be expected to produce docking models within 4Å i-RMSD of the native complex. As the target-template identity increases, so does the expected quality of the final models. For example, most of the 60% identity models produced docking solutions around 2Å i-RMSD. This is likely to represent an overestimate of the achievable quality since one of the docking partners was taken in its bound form. Still it is striking to see that the recent CAPRI targets, which were all homology-homology or homology-unbound docking cases, nicely follow the trend line of our model. This would indicate that the achievable docking quality is limited by the lowest sequence identity component of the interaction partners – in other words: the worse approximation defines the limits of the model.

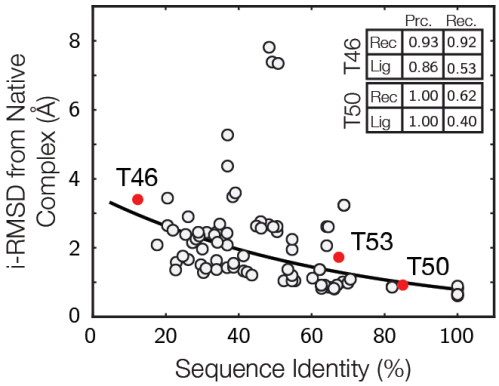


Figure 3. Relationship between sequence identity between target and template with i-RMSD from the native complex of the docking models obtained using CAPRI restraints. Recent CAPRI targets (in black and labeled) follow the observed trend line from the homology-based docking study in this work. Inset Table: Precision (P) and recall (R) of the information gathered for each target. Information for T53 is not yet publicly available. The CAPRI organizing committee communicated the values together with the assessment of the models.

The reliability of the information is of course hard to estimate. During a CAPRI round, most of the information is gathered from literature databases and bioinformatics predictions in the 24-hour period that comprises the server submission. All in all, this essentially means that reliable information is not as scarce as one might imagine. Finally, the homology modelling approach used in this study does not make use of advanced modeling and refinement methods such as those available in structure prediction servers[115,144]. As such, the presented results can be considered a baseline, which can be further improved by expert knowledge of the system under study and/or more powerful structure prediction methods.

Conclusion

Information-driven docking is one of the most reliable and accurate prediction methods for modelling biomolecular complexes. Yet, it needs structural information of the interacting partners as starting point. Given the easy learning curve and widespread availability of homology modelling methods, experimentalists are bound to use them in docking in the absence of experimental alternatives. We have shown here that the global sequence identity between the target sequence and the template used for the modelling of the 3D structure is predictive of the achievable quality of the docking models. Nevertheless, the quality of the

information used to drive the docking remains highly relevant and plays an important role in the outcome of the docking predictions. For templates well inside the so-called 'twilight zone' of sequence identity (~30%), good interface information is sufficient to produce models within 4Å interface backbone RMSD. In contrast, bad quality information can severely diminish the success rate of the docking, even with models built with up to 60% identity. These results allow assessing the suitability of a homology model in information-driven docking and set the stage for more confident predictions even in scenarios where the identity between the template and the sequence is remote, provided that the information on the interaction is reliable.

Acknowledgements

This work was supported by a VICI grant (no. 700.56.442) from the Dutch Foundation for Scientific Research (NWO) and by a Focus and Massa grant from Utrecht University. The WeNMR project (European FP7 e-Infrastructure grant, contract no. 261572, www.wenmr.eu), supported by the European Grid Initiative (EGI) through the national GRID Initiatives of Belgium, France, Italy, Germany, the Netherlands (via the Dutch BiG Grid project), Portugal, Spain, UK, South Africa, Malaysia, Taiwan and the Latin America GRID infrastructure via the Gisela project is acknowledged for the use of web portals, computing and storage facilities.

Supplementary Information

Homology modelling of interacting partners

Each interacting partner of the protein-protein complexes used in this study was modelled using the following protocol:

1. A sequence similarity search was performed on a database of sequences of proteins with known structure ('PDB') using the PSI-BLAST algorithm[145] with default parameters.
2. All hits with low statistical significance (Expectation (E) value greater than 1^{-3}) *or* with an alignment overlap shorter than 85% were discarded.
3. Sequences with identities to the target sequence in the range of 20-80% were binned in steps of 5%, and one representative sequence from each 'bin' was selected.
4. Each sequence was re-aligned to the target sequence using three different algorithms: DALIlite[146], T-Coffee[147], Praline[148] (all with default parameters).
5. Based on the previously derived alignments, MODELLER 9v8[149] was used to build homology models.
6. The 10 best models were selected according to their DOPE score[150].

HADDOCK protocol and energetics details

HADDOCK[30] incorporates biochemical and/or biophysical information as restraints to drive the modelling (E_{AIR}). The docking protocol consists of three steps: i) rigid-body energy minimization stage, ii) semi-flexible refinement consisting of a simulated annealing optimization in torsion angle space with side-chain and backbone flexibility at the interface, and iii) final refinement in explicit solvent (e.g. TIP3P water). Non-bonded interactions were calculated with the OPLS force field[151] using a 8.5Å cut-off. The electrostatic energy (E_{elec}) was calculated using a shifted Coulomb potential, while the van der Waals energies (E_{vdw}) were calculated with a Lennard-Jones potential, with a switching function between 6.5Å and 8.5Å.

In HADDOCK, experimental or prediction information is incorporated in the form of Ambiguous Interaction Restraints (AIRs), a notion similar to that of ambiguous distance restraints based on NOE data in structure calculation of NMR structures (see e.g. Linge et al.[152]). HADDOCK creates AIRs based on active and passive residues defined by the user. For every active residue, HADDOCK creates one single AIR restraint between that residue and all active and passive residues on the partner molecules. These restraints are incorporated in the energy function through a soft-square harmonic potential (E_{AIR}) that depends on an *effective distance*, which is calculated using the following formula:

$$d_{eff}^{iAB} = \left(\sum_{m_{iA}=1}^{Natoms} \sum_{k=1}^{NresB} \sum_{n_{kB}=1}^{NresB} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-\frac{1}{6}}$$

in which *A* and *B* are molecules, *i* iterates over all distance restraints, N_{atoms} indicates all atoms of a given residue and N_{res} the sum of active and passive residues for a given protein. For two

single atoms, the effective distance is equal to the Cartesian distance, but as the number of atoms grows, the effective distance becomes shorter and shorter due to the larger number of distances taken into consideration. By default, HADDOCK enforces an upper limit to the effective distance of 2Å. In case this distance is exceeded, the AIR energy becomes positive and the active residue experiences an attractive force towards the active and passive residues it is restrained to in the partner molecule. Otherwise, the restraint is satisfied and the AIR energy, and consequently the attractive force, is zero. The exact relation between effective distance and energy is described in Nilges *et al.*[68]. Given the many atom-atom distances contributing inversely to the effective distance, a typical AIR restraint is satisfied, depending on the level of ambiguity, if the residue comes within 3-4 Å of any active or passive residues of the partner molecule. As such, (putative) interface residues are driven into (a surface region on) the partner protein, but not to any specific partner residue.

Restraints used in the docking predictions of CAPRI rounds 22-27

Target 46 Unambiguous interaction restraints derived from inter-domain contacts on a distant structural homologue (1P91[153]). The N6 Adenine Specific Dna Methylase partner in T46 (chain B) structurally aligns to the fragment 67-185 of 1P91, while the last beta-strand of Trm112 Activator Protein (chain A in T46) – residues 115-118 aligns with a small beta-strand in 1P91 (residues 33-40). The restraints were implemented between Cα atoms with distances taken from 1P91 with a lower bound of 2Å. Additional unambiguous restraints were defined to keep geometry of the SAH heterogroup and the zinc coordination sites in Mtq2.

Target 47 Unambiguous interaction restraints derived from the interface residues of structure provided by the CAPRI committee. The restraints were implemented between Cα atoms with lower and upper bounds of 0.25Å.

Targets 48 & 49 The interaction restraints are a combination of CPORT[88] predictions and homology-derived restraints using the PDB entry 2YVJ (NADH-dependent ferredoxin reductase and Rieske-type [2Fe-2S] ferredoxin)[154]. We also included unambiguous distances to maintain the geometry of ferrodoxin ironcoordination site.

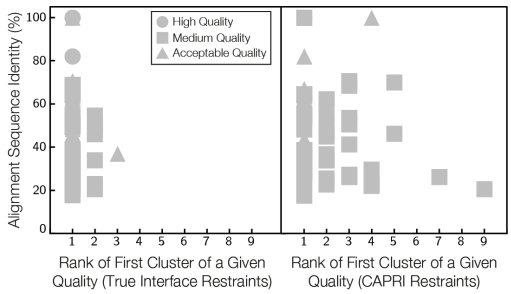
Target 50 The interaction restraints were defined based on a statistical analysis of the most contacted residues on hemagglutinin *(ab initio* docking with 10000 models for it0), together with the assumption that the designed protein HB36.3 was binding to a less mutation-prone surface. On the HB36.3 molecule, the interaction surface was defined based on biophysical properties (aromatic and hydrophobic patches).

Target 51 We only defined connectivity restraints between the different monomers of the complex, together with center of mass (see [88] for details on the implementation) restraints to keep the solutions compact.

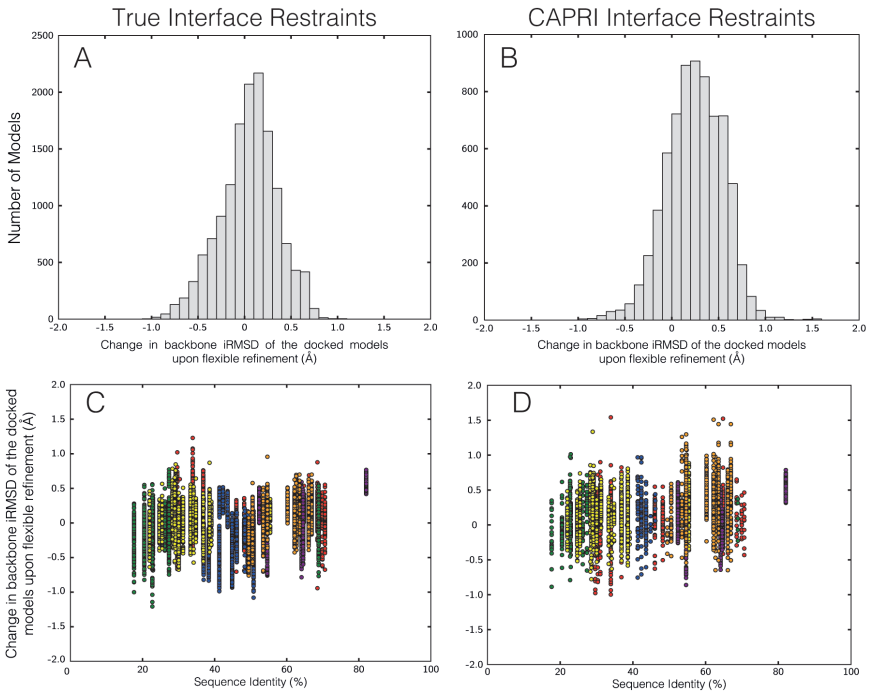
Target 53 & 54 CPORT predictions were used to drive the docking.

Target 57 We defined ambiguous interactions restraints between each sulfate group on the heparin molecule and all arginine and lysine residues (guanidinium and amino groups only) of Bt4661.

Supplementary Figures



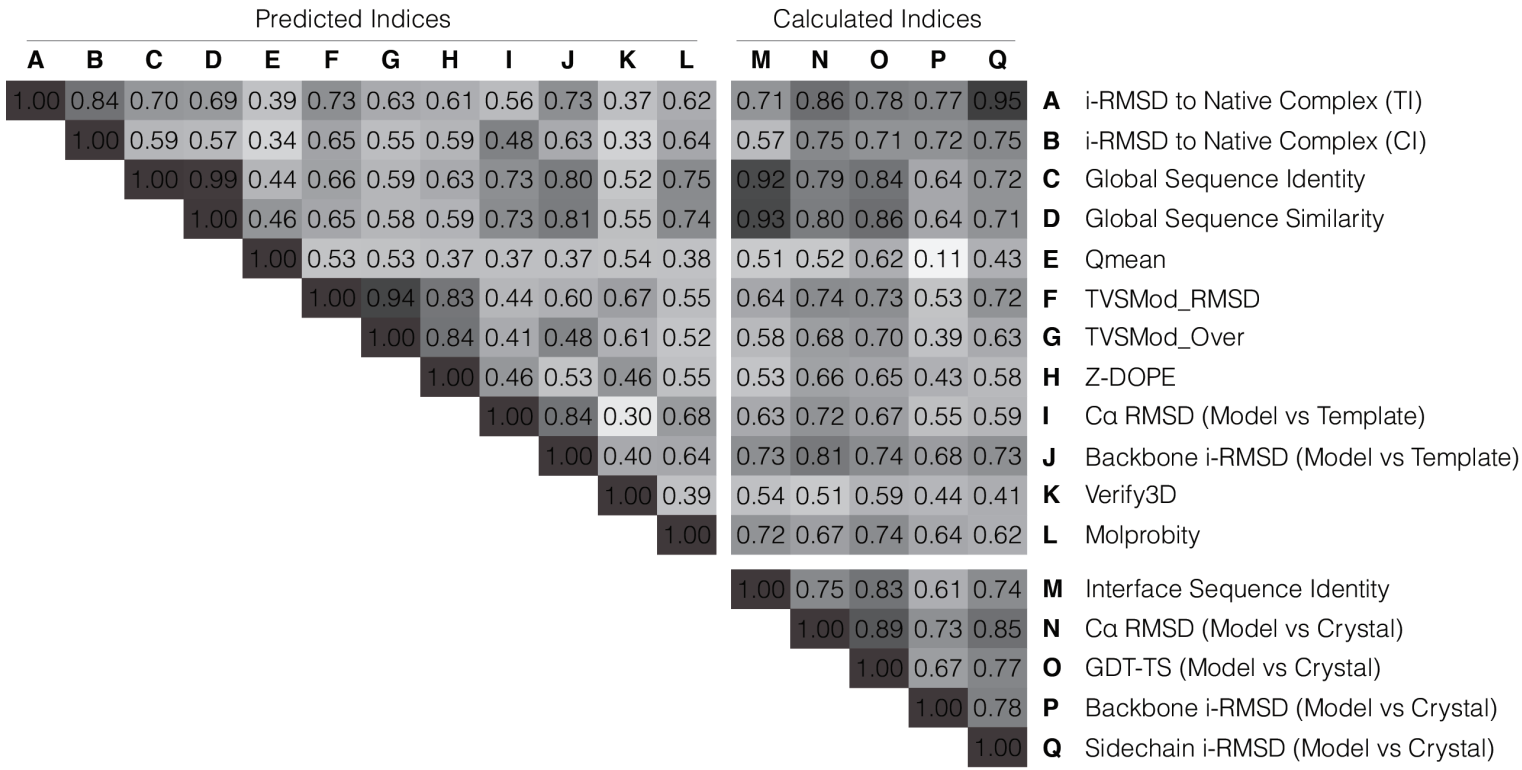
Supplementary Figure 1. Influence of the quality of the homology model and of the data on the scoring performance. The quality of the model is measured by sequence identity between target and template. The impact of the quality of the data is assessed by performing the docking using CAPRI restraints (CI) (right panel) and True interface restraints (TI) (left panel), respectively.



Supplementary Figure 2. Changes in backbone interface RMSD of the predicted complexes upon flexible refinement for all acceptable models (i-RMSD < 4Å) for all targets (21564 models). The top panels **(A & B)** show the distribution of changes as measured by the difference in the backbone interface RMSD of the modelled complex between the rigid-body stage and after final refinement. A negative value means the docked model is closer to the native bound conformation. The bottom panels **(C & D)** show the relation of sequence identity between target/template and change in backbone interface RMSD after flexible refinement with respect to the bound conformation. The models colored according to their CAPRI target: red - T12, blue – T18, green – T26, yellow – T27, purple – T40, orange – T41.

Supplementary Tables

Supplementary Table 1. Cross correlation table of all indices. The values are coded from dark (high correlation) to light (low correlation).



Supplementary Table 2. Details of the methods used for the predictive and measured indices.

Index Name	Type[b]	Calculation Method	Notes
Sequence Identity	P	EMBOSS NEEDLE (EBI)	N/A
Sequence Similarity	P	EMBOSS NEEDLE (EBI)	N/A
Interface Sequence Identity	M	EMBOSS NEEDLE (EBI)	N/A
CαRMSD (model vs. template)	P	ProFit V3.1[a]	Used alignment from homology modelling as ZONEs for fitting.
Backbone interface RMSD (model vs. template)	P	ProFit V3.1[a]	Used alignment from homology modelling as ZONEs for fitting.
MolProbity[143]	P	psvs-1_4-dev.nesg.org	Score is calculated by a log-weighted combination of the clash score, the percentage of Ramachandran Plot not favoured, and the percentage of bad rotamers. Reflects the crystallographic resolution at which these values would be found.
Verify3D[133]	P	psvs-1_4-dev.nesg.org	Score assesses the compatibility between the atomic coordinates of the model and its amino acid sequence, measured by a 3D profile.
Z-DOPE[150]	P	modbase.compbio.ucsf.edu/evaluation	Atomic distance-dependent statistical potential calculated from native structures.
TVSMod_RMSD[131]	P	modbase.compbio.ucsf.edu/evaluation	SVM regression model that predicts the Cα RMSD to the native structure.
TVSMod_Over[131]	P	modbase.compbio.ucsf.edu/evaluation	SVM regression model that predicts the fraction of Cα atoms within 3.5Å of their correct positions in the native structure.
Qmean[132]	P	swissmodel.expasy.org/qmean	Linear combination of six descriptors: two distance-dependent interaction PMF (Cβ atoms and all atoms) to assess long-range interactions; torsion angle potential; solvation potential; agreement with sec. struc. prediction; agreement with BSA prediction.
CαRMSD (model vs. reference)	M	ProFit V3.1[a]	N/A
Backbone Interface RMSD (model vs. reference)	M	ProFit V3.1[a]	N/A
Sidechain Interface RMSD (model vs. reference)	M	ProFit V3.1[a]	Calculated on all atoms except for the backbone, excluding hydrogens.
GDT-TS[155] **(model vs. reference)**	M	as2ts.proteinmodel.org/AS2TS/LGA_list/lga_pdblist.html	Parameters for LGA calculation: -3 –o1 –d:4.0 -gdc

[a] Downloaded at www.bioinf.org.uk/software/profit/ [b] (P)redicted; (M)easured

Supplementary Table 3. Performance of the HADDOCK in the latest CAPRI rounds. Assessment of the best submitted model by interface RMSD, ligand RMSD, Fraction of Native Contacts (FNAT), and Model Quality (CAPRI criteria).

Target	Complex Name	Type	Best Model			
			i-RMSD (Å)	l-RMSD (Å)	FNAT	Model Quality
Manual Submission						
T46	Methyl Transferase Mtq2 /Trm112	HH	3,76	10,1	0,49	Acceptable
T47	Colicin E2 DNase / Im2	UU	0,98	1,52	0,86	High
T48	T4moC / T4moH mono-oxygenase	UU	3,42	9,08	0,23	Acceptable
T49	T4moC / T4moH mono-oxygenase	UU	3,55	14	0,26	Acceptable
T50	HB36.3 designed protein / Flu Hemagglutinin	HU	1,63	6,69	0,67	Medium
T51	Xylanase Xyn10B	UUHU(U)	2,08	6,03	0,41	Acceptable
T53	Designed Rep4 / Rep2 α-repeat	UH	1,73	4,86	0,5	Medium
T54	Designed Neocarzinostatin / Rep16 α-repeat	UH	--	--	--	--
T57	BT4661 / Heparin Complex	UU	2,04	4,68	0,88	Medium
T58	PliG / SalG lysozyme	UU	2,61	6,91	0,29	Acceptable
Server Submission						
T46	Methyl Transferase Mtq2 /Trm112	HH	3,83	11,9	0,57	Acceptable
T47	Colicin E2 DNase / Im2	UU	0,8	1,81	0,86	High
T48	T4moC / T4moH mono-oxygenase	UU	--	--	--	--
T49	T4moC / T4moH mono-oxygenase	UU	3,07	10,7	0,11	Acceptable
T50	HB36.3 designed protein / Flu Hemagglutinin	HU	--	--	--	--
T51	Xylanase Xyn10B	UUHU(U)	--	--	--	--
T53	Designed Rep4 / Rep2 α-repeat	UH	--	--	--	--
T54	Designed Neocarzinostatin / Rep16 α-repeat	UH	--	--	--	--
T57	BT4661 / Heparin Complex	UU	2,11	4,88	0,77	Medium
T58	PliG / SalG lysozyme	UU	--	--	--	--

Chapter 3

Interface prediction of protein-protein complexes from sequence coevolution

T. Hopf*, C.P.I. Schärfe*, J. Rodrigues*, A.G. Green, C. Sander, D.S. Marks, A.M.J.J Bonvin

** These authors contributed equally to this work.*

Adapted from
'Sequence co-evolution gives 3D contacts and structures of protein complexes'
Published in 2014 in Elife, volume 3

Abstract

High-throughput experiments in bacterial and eukaryotic cells identified tens of thousands of interactions between proteins. For a complete understanding, this genome-wide view must be complemented by atomic resolution structures, which unfortunately require low throughput and labor-intensive experiments. Of the many computational alternatives, integrative modeling shows the most promise. Despite being optimal when driven by experimental information, integrative modeling approaches can also make use of interface prediction data. Amongst interface predictors, measures of amino acid coevolution across proteins show great promise. Their application has, however, been limited to specific model systems such as the histidine kinase - response regulator complex. Here we present a new generalized method showing that patterns of evolutionary sequence changes across interacting proteins reflect amino acids that are close in space, with sufficient accuracy to model the three-dimensional structure of the protein complexes. To illustrate the utility of the method, we predict co-evolved contacts between E. coli complexes of unknown structure, including the unknown 3D interactions between subunits of ATP synthase, and produce three-dimensional models consistent with detailed experimental data using HADDOCK. We expect this method to be helpful in cases where experimental data on the interaction is sparse, and as a high-accuracy unambiguous partner-specific interface predictor in cases where sufficient sequence information is available. We also suggest further developments that can possibly open the route for genome-wide applications.

Introduction

A large part of biological research is concerned with the identity, dynamics, and specificity of protein interactions[109]. While there have been impressive advances in the three-dimensional structure determination of protein structures, which has been significantly extended by homology-inferred models, there is still little or no 3D information for~80% of the estimated protein interactions in bacteria, yeast or human organisms[6]. This 'dark matter' of the protein interaction universe amounts to ~30,000/~6000 uncharacterized, but experimentally-verified interactions in human and E. coli, respectively[6]. Since current experimental structure determination methods cannot keep pace with the rapid increase in the demand for residue–level information on these interactions, the scientific community turned to computational modeling methods in order to bridge this knowledge gap.
The most successful modeling methods are hybrid computational-experimental approaches that can combine structural information at varying resolutions, homology models, and other methods, with, residue cross-linking, mutagenesis, and other sources of distance/interface information[13,156]. However, the success of most of these approaches depends on the availability of prior knowledge and many biologically relevant systems, such as membrane proteins, transient interactions, and large molecular complexes remain out of reach, as experimental information is sparse[106].

One promising solution to the lack of residue-level information on protein interactions is to use evolutionary analysis of amino acid co-variation to identify close residue contacts across protein interactions[94,157]. These methods were developed and first used twenty years ago[158,159], though with limited success. More recent approaches have been more successful in identifying residue interactions, as demonstrated for histidine kinases and response regulators[93,160-162], but this approach has yet to be generalized and used to predict contacts between proteins in other protein complexes. In principle, just a small number of key residue-residue contacts across a protein interface would allow computation of reliable structural models and provide a powerful, orthogonal and complementary approach to experiments.

Since the recent and successful demonstration of the use of evolutionary couplings (ECs) between residues to determine the structure of individual proteins[163-166], including integral membrane proteins[167,168], we reasoned that such an evolutionary statistical approach could be used to determine co-evolved residues between proteins (Figure 1A). Upon testing this hypothesis on a dataset of bacterial protein complexes of known structure, we observed that the most correlated residue-residue pairs, according to our approach, were in close spatial proximity in the crystal structure of the complexes. We show here that implementing these pairs as unambiguous distance restraints in HADDOCK[30] successfully generates near-native structural models. Finally, we present the results for evolutionary couplings for a complex of unknown structure, the interaction of subunits a- and b- of *E. coli* ATP synthase, and demonstrate the accuracy of the predictions by comparison with experimental cross-linking data from mass spectrometry.

Material and Methods

Selection of interacting protein pairs for co-evolution calculation.

The candidate set of complexes for benchmark and de novo prediction was derived starting from a dataset of binary protein-protein interactions in E. coli including yeast two-hybrid experiments, literature-curated interactions and 3D complex structures in the

PDB[169]. Since our algorithm for concatenating multiple sequence pairs per species assumes the proximity of the interacting proteins on the respective genomes (see below), we exclude any complex with a gene distance > 10 from further analysis. The gene distance is calculated as the number of genes between the interacting partners based on an ordered list of genes in the E. coli genome obtained from the UniProt database[170]. The resulting list of pairs (~ 300) was then filtered for redundant homologous complexes and heterodimers based on the identification of Pfam[171] domains in the interacting proteins (236). All remaining complexes with a known 3D structure or an interlog 3D structure (71) (identified by intersecting the results of HMMER[172] searches against the RCSB PDB[173] for both monomers) were used for benchmarking the prediction method.

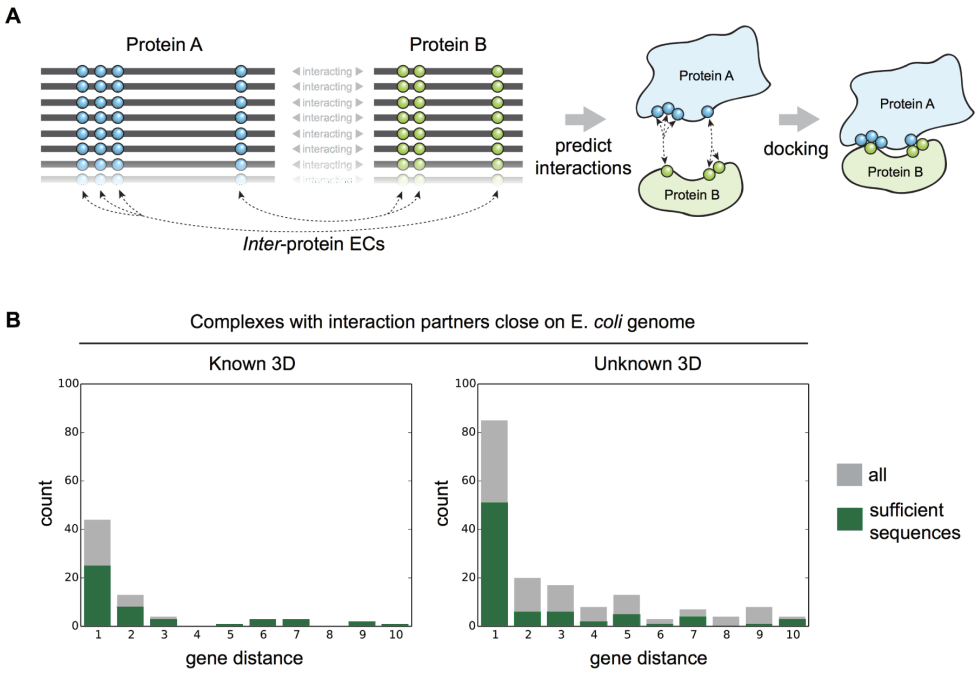


Figure 1. Co-evolution of residues across protein complexes from the evolutionary sequence record. **(A)** Evolutionary pressure to maintain protein-protein interactions leads to the coevolution of residues between interacting proteins in a complex. By analyzing patterns of amino acid co-variation in an alignment of putatively interacting homologous proteins (left), evolutionary couplings between coevolving inter-protein residue pairs can be identified (middle). By defining distance restraints on these pairs, docking software can be used to calculate the 3D structure of the complex (right). **(B)** Distribution of E. coli protein complexes of known and unknown 3D structure where both subunits are close on the bacterial genome (grey bars), allowing sequence pair matching by genomic distance. For a subset of these complexes, sufficient sequence information is available for evolutionary couplings analysis (green bars).

Multiple sequence alignments

Each protein from all pairs in our dataset was used to seed a multiple sequence alignment (MSA) using jackhmmer[174] and the alignment depth for the calculations was chosen to optimize the number of sequences retrieved and the coverage of the protein length, as in previous work[167]. In order to calculate co-evolved residues across different proteins we need to match the pairs of interacting protein sequences. To match the pairs, we assume interacting proteins are located in close proximity on their respective genomes, often on the same operon, as in the methods used previously matching histidine kinase and response regulator interacting pairs[93,161]. We retrieved the genomic locations of proteins in the alignments and concatenated pairs following 2 rules. (i) the CDS of each concatenated protein pair must be located on the same genomic contig (using ENA[175] for mapping), and (ii) each pair must be the closest to one another (on genome), when compared to all other possible pairs. The concatenated sequence pairs were filtered based on the distribution of genomic distances to exclude outlier pairs with high genomic distance (Figure 1). Alignment members were clustered together at 80% sequence identity and afterwards reweighted, thus implicitly removing duplicate sequences from the alignment. Supplementary Table 1 reports the total number of concatenated sequences, the lengths, and the effective number of sequences remaining after down-weighting highly similar proteins and removing fragments as has been done in previous work[164].

Evolutionary couplings calculation

Inter- and intra-ECs were calculated on the alignment of concatenated sequences using a global probability model of sequence co-evolution, adapted from the method for single proteins[163,164,167] using a PLM[176] rather than mean field approximation to calculate the coupling parameters. Columns in the alignment that contain more than 80% gaps were excluded and the weight of each sequence was adjusted to represent its cluster size in the alignment thus reducing the influence of identical or near-identical sequences in the calculation. We can then compare the predicted ECs for both within and between the protein/domains to the crystal structures of the complexes.

To predict the accuracy of the calculated inter-EC, we examined the rank of the first inter-EC for each complex, relative to all intra- and inter-ECs, calculated on the concatenated alignment. For instance the length of the concatenated MoaD/MoaE alignment is 231, resulting in 26,565 pairs of EC scores, of which all 12150 inter-protein EC scores and a subset of 13290 intra-EC scores remain when excluding intra-EC pairs with a primary sequence distance of up to 5. The highest rank inter-EC is 11th in the combined inter and intra EC list. Using these observations we defined an inter-EC score based on the relative rank of the first inter-protein residue pair in the list of all intra and inter pairs of ECs calculated from the multiple sequence alignment. Based on previous work on single proteins, we filter for sequence sufficiency requiring > 0.5 sequences per residue in the concatenated sequence by counting the effective number of sequences after redundancy (Table 1).

Docking

Monomer structures for each of the proteins in the HK-RR and CLPS-CLPA complexes and ATPE were taken from crystallized unbound conformations. For the other benchmark complexes we randomized the side chains of the monomers before docking because subunits that have been crystallized together in a complex will be biased due to the complementary

positions of the surface side chains, and hence docking these proteins, even without distance restraints, will produce high-ranking correct structures. Therefore, starting monomers (i.e. those extracted from complex structures) were subjected to side chain replacement using SCWRL4[177] resulting in ~1.5Å RMSD over the side chain atoms relative to the original structure. We used HADDOCK[30], a widely used docking program based on ARIA[152] and the CNS (Crystallography and NMR System) software[64], to dock the monomers for each protein pair with 5, 10 inter-ECs as unambiguous distance restraints on the Cα atoms of the backbone.

Each docking calculation starts with a rigid-body energy minimization, followed by semi-flexible refinement in torsion angle space, and ends with further refinement of the models in explicit solvent (water). 500/100/100 models generated for each of the three steps, respectively. All other parameters were left as the default values in the HADDOCK protocol. Each protein complex was run using predicted ECs as unambiguous distance restraints on the Cα atoms (with an upper bound of 7Å and a lower bound 3Å; input files available upon request). As a negative control, each protein complex was also docked using center of mass restraints (*ab initio* docking mode of HADDOCK)[88] alone and in the case of the controls generating 10000/500/500 models. Each of the generated models is scored using a weighted sum of electrostatic (Eelec) and van der Waals (Evdw) energies complemented by an empirical desolvation energy term (Edesolv)[129]. The restraints energy term was explicitly removed from the scoring function in the last iteration (Edist3 = 0.0) to enable comparison of the scores between the runs that used a different number of ECs as distance restraints:

$$\text{HADDOCK Score} = 0.2 \times E_{elec} + 1.0 \times E_{vdW} + 1.0 \times E_{desolv}$$

Comparison of models to crystal structures

All models in the benchmark were compared to the cognate crystal structures by the root mean square deviation (RMSD) of all backbone atoms at the interface of the complex using ProFit v.3.1 (http://www.bioinf.org.uk/software/profit/). The interface was defined as the set of all residues that contain an atom within 6Å from any atom of the complex partner. For the ATPE-ATPG complex we excluded the 2 C-terminal helices of ATPE as these helices are mobile and take many different positions relative to other ATP synthase subunits[178]. Similarly, since the DHp domain of histidine kinases can take different positions relative to the CA domain, the HK-RR complex was compared over the interface between the DHp domain alone and the response regulator partner. Accuracy of the models generated with EC restraints was compared with that of the models with center of mass restraints alone (negative controls) (Supplementary Table 2).

Results

Methodological approach

We investigated whether coevolving residues between proteins are close in three dimensions by assessing blind predictions against experimentally determined structures of known protein complexes. Beginning with a published dataset of ~3500 high-confidence protein interactions in E. coli[169], we removed redundancy and conditioned their inclusion based on genome distance between the pairs of proteins in the *E. coli* genome resulting in 236 interactions, see Materials and Methods. Our current algorithm leverages the fact that prokaryote genomes contain particularly organized structures, the operons, which encode

functionally related genes in close proximity under the same promoter[179], or that there are only single occurrences of the pair of interacting proteins in each genome (Table 1).

Before calculating the evolutionary couplings, the sequences of each complex were concatenated and columns in the resulting paired alignment containing more than 80% gaps removed. To reduce the influence of identical or near-identical sequences in the calculation, we clustered similar sequences at 80% sequence identity) and weighted them to represent their cluster size in the alignment. The residue-residue correlations were then calculated using EVcouplings[163,164,180], which implements a pseudo-likelihood maximization (PLM) approximation to determine the interaction parameters in the maximum entropy equation[166,180]. This approach generates both intra- and inter-EC scores for all pairs of residues within and across the protein pairs (Figure 1A). Benchmark calculations here and in previous work[163,164] indicate that the number of sequences in the alignment is critical (at least 0.5 non-redundant sequences per residue, see also Materials and methods) and that the accuracy of ECs that rank lower than L (the number of residues in the sequence) decays rapidly. Based on these observations we use the strength of the top inter-protein EC (relative to all intra-protein EC residue pairs) to estimate the likelihood of the predicted interactions to be accurate. We compared our predicted inter-protein ECs against residue distances in a subset of seven known crystal structures of complexes that had sufficient sequences together with high inter- EC scores (< 0.3).

The top-ranked ECs are mostly inter-protein contacts

For the top ranked benchmark complexes regarding the interEC score, the majority of the top 5 ECs between proteins is correct to within 8Å (Table 1 and Supplementary Table 2). The normalized inter EC-score demarks those complexes that have accurate EC predictions; inter-EC scores < 0.3 distinguishes mostly accurate ECs in the top 5 predictions (Table 1, Figure 3), while those with inter-EC scores > 0.6 have mostly inaccurate ECs (data not shown). Since the coevolution score indicates incorrect ECs, we reasoned that the scores could also distinguish interaction from non-interaction of protein pairs.

We tested this on the a-, b-, c- and α-subunit interactions of ATP synthase and correctly paired the interacting subunits using the inter-protein EC scores (Figure 4A and B). Some of the high confidence predictions are not spatially close when compared to crystal structures. These false positives may be a result of a number of reasons, including incorrect assumptions about interacting pairs across the alignments and oligomerization signals. The complexes may also exist in alternative conformations not necessarily captured by a single crystal structure, for instance in the case of the large conformational changes of the BtuCDF complex[181]. To test whether the inter-protein ECs are sufficient for computing accurate structural models, we selected seven examples for docking. Using high ranked interEC pairs (5 and 10) as distance restraints between the interacting proteins we generated models of the complexes using HADDOCK[30]. Over 70% of the generated models were close to the crystal structures of the complexes, as measured by the interface backbone root mean square deviation (i-RMSD) of atomic coordinates (Figure 2, Supplementary Table 2).

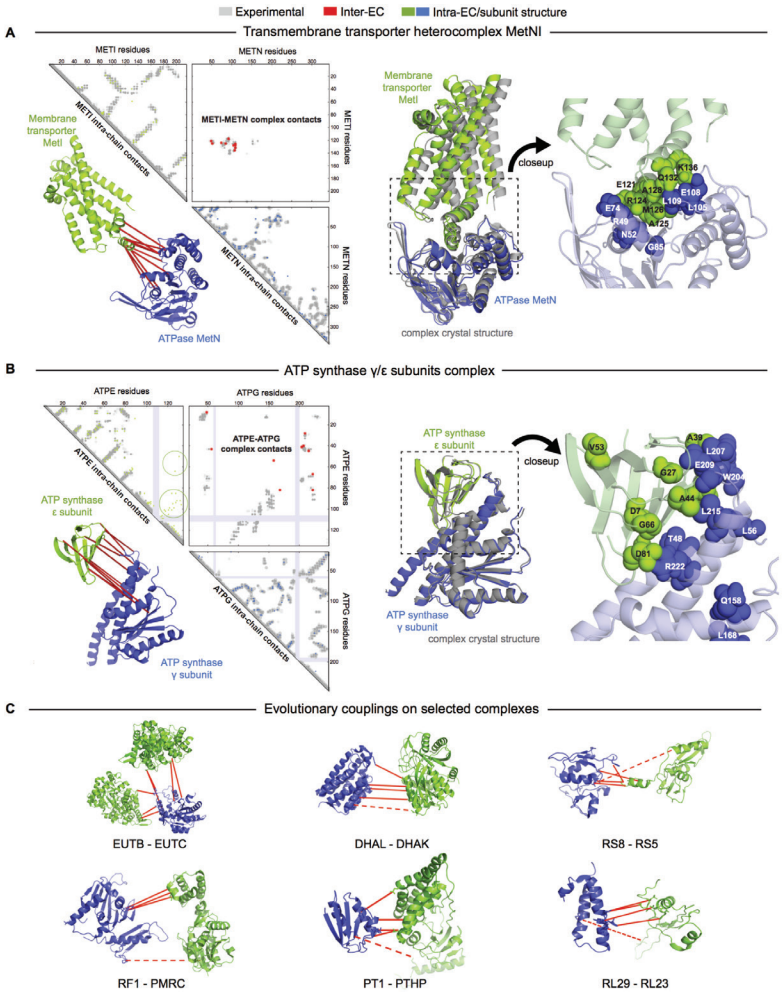


Figure 2. Evolutionary couplings give accurate 3D structures of complexes. EVcomplex predictions and comparison to crystal structure for **(A)** the methionine-importing transmembrane transporter hetero-complex MetNI from E. coli (PDB: 3tui) and **(B)** the gamma/epsilon subunit interaction of E. coli ATP synthase (PDB: 1FS0). **(A,B)** Left panels: Complex contact map comparing predicted top ranked inter-ECs (red stars, upper right quadrant) and intra-ECs (to the occurrence of the 10th inter-EC; green and blue stars, top left and lower right triangles) to close pairs in the complex crystal (dark/mid/light grey points for minimum atom distance cutoffs of 5/6/7 Å; missing crystal data: shaded blue rectangles). The top 10 inter-ECs are also displayed on the spatially separated subunits of the complex (red lines on green and blue cartoons, lower left). Right panels: Superimposition of the top ranked model from 3D docking (green/blue cartoon, left) onto the complex crystal structure (grey cartoon), and close-up of the interface region with highly coupled residues (green/blue spheres, residues involved in top 10 inter-ECs). **(C)** High-ranking inter-ECs on a selection of benchmark complexes (hetero-complex subunits: blue, green cartoons, multiple copies of the same subunit in identical color; correct inter-ECs: solid red lines, incorrect inter-ECs: dashed red lines).

Evolutionary conserved residue networks across proteins

The top 10 inter-EC pairs between MetI and MetN are accurate to within 8Å in the MetNI complex (PDB: 3TUI [182]), resulting in an average - of 1.6±0.6Å from the crystal structure for all generated HADDOCK models (Table 1). The top three inter-EC residue pairs (K136-E108, A128-L105, and E121-R49, MetI-MetN respectively) constitute a residue network coupling the ATP binding pocket of MetN to the membrane transporter MetI. This network calculated from the alignment corresponds to residues identified experimentally that couple ATP hydrolysis to the open and closed conformations of the MetI dimer[182] (Figure 2A). The vitamin B12 transporter (BtuC) belongs to a different structural class of ABC transporters, but also uses ATP hydrolysis via an interacting ATPase (BtuD). Our approach identifies sufficient interprotein residue contacts to calculate accurate models with HADDOCK. The top ranked model is 0.8Å backbone i-RMSD from the crystal structure, 1L7V[183], (Table 1, Supplementary Table 2). The top five inter-ECs co-locate the L-loop of BtuC close to the Q-loop ATP-binding domain of the ATPase, hence coupling the transporter with the ATP hydrolysis state in an analogous way to MetI-MetN. The identification of these coupled residues across the different subunits suggests that our approach identifies not only residues close in space, but also particular pairs that are constrained by the transporter function of these complexes[182,184], reflecting the original assumption that structurally and functionally important amino acid pairs are maintained throughout evolution.

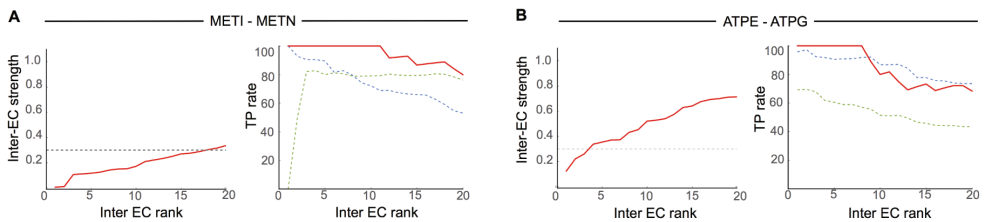


Figure 3. The accuracy of inter-protein ECs is predicted by their relative rank. (A) Evolutionary couplings between MetI and MetN are high relative to all intra- and inter-protein ECs (left plot, red). 16 inter-ECs have a normalized inter-EC score <0.3, of which 85% are true positives (right plot, red line) (B) In contrast, compared to MetI-MetN, fewer ECs between AtpE and AtpG have high inter-EC scores (left plot, red line), but all 4 are true positives (right plot, red line). The normalized inter-EC strength is the rank percentile of the inter-EC in the list of top L intra- and inter- ECs, where L is the concatenated length of the proteins in the complex. (A, B) The true positive rates for the inter-protein ECs weakly corresponds the accuracy of the intra-protein ECs for the individual proteins (right plots, green and blue dashed lines).

The ATP synthase ε and γ subunit complex provides a challenge to our approach, since the ε subunit can take different positions relative to the γ subunit, executing the auto-inhibition of the enzyme by dramatic conformational changes[178]. In a real- world scenario, where we might not know this a priori, there may be conflicting constraints in the evolutionary

record corresponding to the different positions of the flexible portion of ε subunit. Our approach accurately predicts 8 of the top 10 inter-EC pairs (within 8Å in the crystal structure 1FS0[185], with the top inter-EC, εA40-γL207, bridging the subunits at the end of an inter-protein beta sheet. The C-terminal helices of the ε subunit are significantly different across 3 crystal structures (PDB IDs: 1FS0, 1AQT[186], 3OAA[178]). The top ranked intra-ECs support the conformation seen in 1aqt, with the C-terminal helices tucked against the N-terminal beta barrel (Figure 2B, green circles) and do not contain a high ranked evolutionary trace for the extended helical contact to the γ subunit seen in 1FS0 or 3OAA. Docking with the top inter-ECs results in models with 1.5/1.3Å backbone i-RMSD from the crystal structure, for the interface between the N-terminal domain of the ε subunit and the γ subunit (Table 1, Figure 2B). εD81 and γR222 connect the ε- subunit via a network of three high intra ECs between the N and C terminal helices to the core of the F1 ATP synthase. In summary, these benchmarks show the power of evolutionary information to infer protein complex structure phenotypes, and indicate criteria that can estimate the success of the predictions, particularly valuable in case of complexes of unknown structure. The benchmarks also show that ECs provide precise relationships across the proteins that could be critical for the identification of functional coupling pathways in addition to the 3D models.

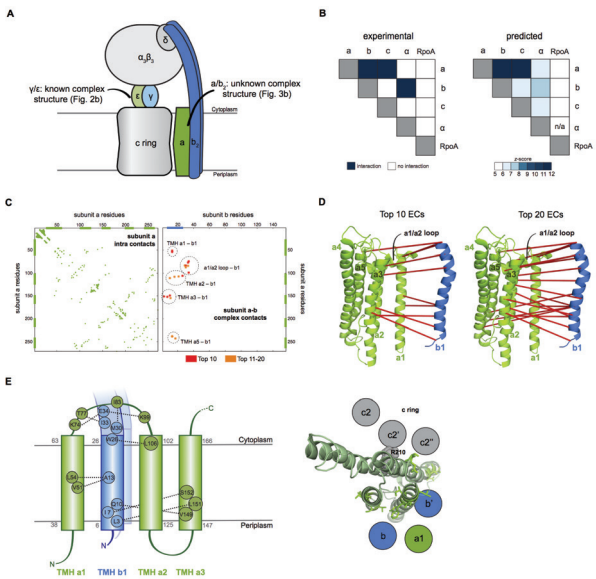


Figure 4. Predicted interactions between the a-, b- and c- subunits of ATP synthase. **(A)** The a- and b-subunits of E. coli ATP synthase interact, but the monomer structure of both the individual subunits and the complex is unknown. **(B)** Experimental evidence for binary interactions between the ATP synthase subunits a, b, c, and α and RNA polymerase subunit α (negative control) agrees with the predictions. **(C)** Complex contact map of 10 top ranked inter-ECs and corresponding top ranked subunit a intra-ECs (subunit b intra-ECs not shown). **(D)** Inter-ECs (red lines) between subunit a (model predicted with EVfold-membrane, green) and subunit b (PDB: 1b9u, blue). **(E)** Left panel: Residue detail of predicted interaction between subunit a and b (dotted lines, predicted transmembrane helices as grey numbers). Right panel: Proposed helix-helix interactions between ATP synthase subunits a (green), b (blue, homodimer), and the c ring (grey).

Table 1. Accuracy of coevolution-based prediction of residue-residue contacts.

Complex[a]	PDB ID	L Ratio[c]	Relative interEC rank[d]	False Positive ECs[e]	i-RMSD (Å)[f]	
					Top Ranked Model	Best Model
ATP synthase γ and ε subunits[b]	1FS0	1.9	0.12	0/2	1.3/1.5	1.4/1.3
Vitamin B12 uptake system permease & ATP-binding domain	1L7V	4.4	0.02	1/1	0.8/1,5	0.8/0.8
Vitamin B12 uptake system SBP & permease	2QI9	3.8	0.17	1/5	2.5/1.4	2.1/1.4
Methionine Transporter Complex	3TUI [187]	1.1	0.02	0/0	1.9/1.6	1.7/1.5
Molybdopterin synthase	1FM0 [188]	2.0	0.05	0/0	5.5/2.9	5.2/2.9
Histidine Kinase - Response Regulator complex[b]	3DGE	38.4	0.09	0/0	1.7/1.9	1.6/1.7
ClpAS chaperone-protease complex[b]	1R6Q	0.7	0.03	1/5	2.0/3.3	2.0/2.0

[a] UniProt IDs for complex pairs: ATPE_ECOLI/ATPG_ECOLI, BTUC_ECOLI/BTUD_ECOLI, BTUC_ECOLI/BTUF_ECOLI, METI_ECOLI/METN_ECOLI, MOAD_ECOLI/MOAE_ECOLI, Q9WZV7_THEMA/Q9WYT9_THEMA, CLPS_ECOLI/CLPA_ECOLI FENR_ANASO/FER1_ANASO, TRXB_ECOLI/THIO_ECOLI, HIS6_THEMA/HIS5_THEMA

[b] Docking started from unbound (free) structures of the interacting partners.

[c] Reweighted number of sequences per residue of concatenated length

[d] Rank of first interEC normalized by the concatenated alignment length.

[e] False positive contacts (>8Å in crystal structure) in top 5/10 ECs.

[f] Values provided for the two docking calculations with the top 5 and top 10 ECs.

De novo predictions of the interactions between the a- b- c- subunits of ATP synthase

The structure of the a-, b- and c-subunit interaction of ATP synthase is of wide biological interest (reviewed in Walker et al.[189]), and has proven experimentally challenging[190] (Figure 4A). Since the structure of the membrane-integral penta-helical a-subunit is unknown, we generated a *de novo* model using evolutionary couplings as described previously[167]. The resulting model of the a-subunit is consistent with topologies that have been inferred from mass spectrometry crosslinking studies[191-193]. The ten highest-ranked inter-ECs between the a- and b- subunits (Figure 4C, D and E) are all consistent with experimental cross-linking studies (Supplementary Table 3) and the proposed geometry of the interaction[194] (Figure 4E). The top ranked EC between the a- and b-subunit (aK74 - bE34) coincides with experimental evidence of the interaction of aK74 with the b-subunit[191,192].

Most other high-ranking ECs involve the cytoplasmic loop connecting transmembrane helices TMHa1 and TMHa2 in the a-subunit, with residues 30-34 in the b- subunit (Figure 4C, D and E). The top EC between the a- and c-subunits (aG213 cM64) lies close to the functionally critical aR210, cD61 interaction[195] on the same helical faces of the respective subunit (Figure 4E).

In general, the agreement between our de novo predicted inter-protein ECs with available experimental data serves as a measure of confidence for the predicted residue pair interac-

tions, and suggests that our approach can be used to reveal structural details of yet unsolved protein complexes given sufficient evolutionary information.

Discussion

A primary limitation of our approach is its dependence on the availability of a large number of evolutionarily related sequences. For proteins in a complex, each one hundred residues in length, we estimate that one needs more than 100 sequences for the concatenated proteins, after filtering for redundancy and down- weighting highly similar proteins in the alignments, i.e. 0.5 sequences per residue in the complex (Table 1). Although there are nearly 3000 sequenced bacterial genomes, this number does not provide the necessary coverage for many protein complexes. If there are multiple orthologs of the interacting proteins, as in the case of the HK-RR interaction, then our method can successfully pair proteins and hence construct correctly concatenated alignments. This limitation is more stringent for eukaryotic protein complexes, given their more recent evolutionary divergence. However, with the rapidly expanding number of sequenced genomes it is plausible that we will be able to explore far higher percentage of protein interactions in the near future. A second limitation is the dependence of the approach on genome proximity of the interacting pairs of proteins, which is required in order to distinguish interacting from non-interacting paralogs. Since only ~10% of interacting protein pairs in *E. coli* are distributed in this way, the challenge is now to develop an algorithm that can calculate the probability that any pair is co-evolved.

Conclusion

The work presented here is in anticipation of the above-mentioned genome-wide exploration and, as a proof of principle, shows the accurate prediction of inter-protein contacts and their ability to predict structural models of diverse protein-protein complexes. As with single protein (intra-EC) predictions, evolutionary conserved conformational flexibility and oligomerization can result in more than one set of contacts that must be de-convoluted. Can evolutionary information help to predict the details and extent for each complex? The challenge will certainly involve the development of algorithms that can disentangle evolutionary signals caused by alternative conformations of single complexes, alternative conformations of homologous complexes, or simply false positive signals. Taken together, these limitations highlight fruitful areas for future development of the presented methodology. Despite strong requirements for the successful de novo calculation of co-evolved residues, the power of the method illustrated here may hugely accelerate the exploration of the protein- protein interaction world and the determination of protein complexes on a genome-wide scale. The use of co-evolutionary analysis towards computational models to determine protein specificity and promiscuity, co-evolutionary dynamics and functional drift will be exciting future research questions.

Supplementary Tables

Supplementary Table 1. Alignment Statistics for the Benchmark Complexes

Complex	Uniprot Id		E-value cutoff		Number of Sequences				Concatenated L	Rank of First interEC	inter-EC score
	A	B	A	B	A	B	A+B	A+B (filtered)			
ATP synthase γ and ε subunits	ATPE_ECOLI	ATPG_ECOLI	2	2	10807	10236	8619	802	426	206	0,12
Vitamin B12 uptake system permease & ATPbinding domain	BTUC_ECOLI	BTUD_ECOLI	10	51	57284	55663	19927	2533	575	61	0,03
Vitamin B12 uptake system SBP & permease	BTUC_ECOLI	BTUF_ECOLI	10	5	57284	33657	14873	2272	592	399	0,17
Methionine Transporter Complex	METI_ECOLI	METN_ECOLI	23	78	35145	85510	18115	625	560	46	0,02
Molybdopterin synthase	MOAD_ECOLI	MOAE_ECOLI	1	1	7657	6745	5217	462	231	67	0,05
Histidine Kinase / Response Regulator complex	Q9WZV7_THEMA	Q9WYT9_THEMA	25	25	180343	149862	78510	14216	370	34	0,09
ClpAS chaperone-protease complex	CLPS_ECOLI	CLPA_ECOLI	1	10	5978	22526	4744	161	247	41	0,03

Supplementary Table 2. Docking Results for the Benchmark Complexes

Complex	PDB Id		False Positives[a]		Negative Control		Best 5 ECs		Best 10 ECs	
	A	B	Top 5 ECs	Top 10 ECs	Top Ranked iRMSD (Å)	No. Acceptable Models[b]	Top Ranked iRMSD (Å)	No. Acceptable Models[b]	Top Ranked iRMSD (Å)	No. Acceptable Models[b]
ATP synthase γ and ε subunits	1AQT	1FS0:G	0	2	9,5	0	1,5	100	1,3	100
Vitamin B12 uptake system permease & ATPbinding domain	1L7V:A	1L7V:C	1	1	16,3	0	0,8	100	1,5	100
Vitamin B12 uptake system SBP & permease	2QI9:AB	2QI9:F	1	5	24,1	0	2,5	32	1,4	100
Methionine Transporter Complex	3TUI:E	3TUI:G	0	0	38,7	0	1,9	100	1,6	100
Molybdopterin synthase	1FM0:D	1FM0:E	0	0	18,6	0	5,5	0	2,9	40
Histidine Kinase / Response Regulator complex	3DGE:A	3DGE:C	0	0	4,6	6	1,7	100	1,9	100
ClpAS chaperone-protease complex	1R6Q:C	1R6Q:A	1	5	8,7	7	2,0	81	3,3	100

[a] A predicted pair further than 8Å in the crystal structure (minimum atom distance) is considered a false positive.
[b] Acceptable quality if defined as i-RMSD smaller than 4Å from the crystal structure of the complex. No. of models out of 100.

Supplementary Table 3. Experimental data for a- and b- subunit interactions of ATP synthase

EC pair Rank	Position in subunit a	Subunit a	Subunit b	Experimental validation	Reference(s)
1	Cytoplasmic loop TMH1-TMH2	74K	34E	Crosslinking direct	Long, J. C., et al., Characterization of the first cytoplasmic loop of subunit a of the Escherichia coli ATP synthase by surface labeling, cross-linking, and mutagenesis. J Biol Chem 277, 27288–27293 (2002) Caviston, T. et al., Identification of an uncoupling mutation affecting the b subunit of F1F0 ATP synthase in Escherichia coli. FEBS Lett 429, 201–206 (1998).
2	Cytoplasmic loop TMH1-TMH2	77T	33I	Crosslinking neighborhood	
9	Cytoplasmic loop TMH1-TMH2	83I	30M	Crosslinking neighborhood	
18	Cytoplasmic loop TMH1-TMH2	83I	29L	Crosslinking neighborhood	
7	Cytoplasmic loop TMH1-TMH2	99K	34E	Crosslinking neighborhood	
11	Cytoplasmic loop TMH1-TMH2	87N	30M	Crosslinking neighborhood	
19	Cytoplasmic loop TMH1-TMH2	84G	33I	Crosslinking neighborhood	
8	TMH2	106L	26W	Crosslinking neighborhood	Long, J. C., et al., Characterization of the first cytoplasmic loop of subunit a of the Escherichia coli ATP synthase by surface labeling, cross-linking, and mutagenesis. J Biol Chem 277, 27288–27293 (2002) DeLeon-Rangel, J., Ishmukhametov, R. R., Jiang, W., Fillingame, R. H. & Vik, S. B. Interactions between subunits a and b in the rotary ATP synthase as determined by cross-linking. FEBS Lett 587, 892–897 (2013).
14	TMH2	111W	10Q	Crosslinking neighborhood	
15	TMH2	108I	16L	Crosslinking neighborhood	
16	TMH2	107T	21C	Crosslinking neighborhood	
10	TMH1 middle	51V	13A	Not tested	-
5	TMH1 middle	54L	13A	Not tested	
3	TMH3	149V	10Q	Crosslinking neighborhood	DeLeon-Rangel, J., et al.,. Interactions between subunits a and b in the rotary ATP synthase as determined by cross-linking. FEBS Lett 587, 892–897 (2013).
4	TMH3	152S	7I	Crosslinking neighborhood	
6	TMH3	151L	3L	Crosslinking neighborhood	
20	TMH3	155L	11A	Crosslinking neighborhood	
12	Periplasmic end of TMH5	238N	13A	Not tested -but a227/8 contacts bN2	-
17	Periplasmic end of TMH5	239V	13A	Not tested -but a227/8 contacts bN2	
13	Periplasmic end of TMH5	243I	17F	Not tested -but a227/8 contacts bN2	

Chapter 4

Clustering biomolecular complexes by residue contacts similarity

J. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca,
A. Melquiond, A.M.J.J Bonvin

Published in 2012 in Proteins, volume 80, issue 7

Abstract

Inaccuracies in computational molecular modelling methods are often counterweighed by brute-force generation of a plethora of putative solutions. These are then typically sieved via structural clustering based on similarity measures such as the root mean square deviation of atomic positions. Albeit widely used, these measures suffer from several theoretical and technical limitations (e.g. choice of regions for fitting) that impair their application in multi-component systems (N>2), large-scale studies (e.g. interactomes), and other time-critical scenarios. We present here a simple similarity measure for structural clustering based on atomic contacts – the fraction of common contacts – and compare it with the most used similarity measure of the protein docking community – interface backbone RMSD. We show that this method produces very compact clusters in remarkably short time when applied to a collection of binary and multi-component protein-protein and protein-DNA complexes. Furthermore, it allows easy clustering of similar conformations of multi-component symmetrical assemblies in which chain permutations can occur. Simple contact-based metrics should be applicable to other structural biology clustering problems, in particular for time-critical or largescale endeavors.

Introduction

The road to complete comprehension of a biological process inevitably passes through the knowledge of the detailed atomic structures of its participants[197]. Unfortunately, experimental determination of biomolecular structures is often problematic and time consuming, while *in silico* molecular modelling methods devised as complementary approaches suffer from chronic inaccuracy due to the simplified physics they are based on[198]. Nevertheless, the relative ease and speed with which the latter yield near-atomic resolution models earned them a spot in the limelight of structural biology methods.

To counterweigh their innate inaccuracy, molecular modelling methods often generate thousands to tens of thousands of possible conformations for a single structure, each representing a discrete point in its energy landscape. A posterior selection process is then necessary to salvage the most native-like conformations. Previous research has shown that, since a native structure is very unlikely to be an isolated event in the energy landscape, it is expected to neighbor similar near-native conformations in a basin with overall low potential energy[52]. This observation hinted at the adoption of clustering techniques, devised to group elements sharing common attributes, to the benefit of the selection process. In fact, it has been shown in both protein structure prediction[52] and protein-protein docking[199] that clustering indeed helps discriminate near-native structures better than energetics alone. Predictably, the most successful algorithms at both CASP (Critical Assessment of Techniques for Protein Structure Prediction)[200] and CAPRI (Critical Assessment of Prediction of Interactions)[9] experiments have incorporated at least one clustering step in their protocols.

The performance of clustering algorithms is nevertheless dependent on the similarity measure used to determine the similarity between any two elements of a dataset, which for the majority of the state-of-the-art clustering algorithms is the root mean square deviation (RMSD) of atomic coordinates. Yet, previous research has shown that, despite widely adopted, RMSD suffers from several shortcomings. First, it loses sensitivity as the molecular weight of the system increases, since large regions with little deviations become dominant[201]. Second, and more importantly, the necessity of choosing the regions to fit the structures under scrutiny to one another results in biased measurements. Finally, RMSD calculations are CPU intensive and consume a large amount of live memory (RAM), yet another hindrance to the structural comparison of increasingly larger and more complex systems.

This conjecture motivated several studies comparing and assessing similarity measures[202,203]. Metrics such as dihedral angles and variants of RMSD such as distance matrix RMSD have been used to cluster molecular dynamics trajectories. It has also been shown that a metric based on residue contacts – contact matrix distance – accounted for less chaotic clusters. Furthermore, contact-based measures (lDDT[204] and FNAT[9]) are already being used to assess the quality of submitted models in CASP and CAPRI, respectively.

The protein docking community is shifting its focus to more intricate systems such as entire interactomes or supramolecular assemblies consisting of a large number of components[106]. As a result, similarity measures that retain a high sensitivity while performing substantially faster than traditional RMSD-based metrics are required. Inspired by the widespread usage of contact information in structure comparison, we theorized that calculating the fraction of common contacts (FCC) between two structures, akin to the notion of fraction of native contacts used in CAPRI, would first, describe the relative orientation of the interacting partners, and second, provide detailed residue-level information. Such

a measure, if applied to structural clustering, should yield sufficient discriminatory power without suffering from any of the theoretical downsides of positional RMSD measures, and save a considerable amount of computation time since it discards the structural alignment step. In the following, we introduce the concept of FCC clustering and demonstrate its performance in a set of binary and multimeric complexes, selected not only to reflect typical scenarios in protein docking but also challenging cases including assemblies with internal symmetry and protein-DNA complexes.

Materials and Methods

Identifying Residue Contacts

For each non-hydrogen atom pair (*i, j*) in a structure, the Euclidean distance between the atoms () is computed. If this distance is below a threshold and both atoms belong to different polypeptide chains, the pair of residues to which the atoms belong to is considered to be in contact. We defined as 5Å in accordance with CAPRI criteria, the standard in the docking field.

Calculating the Fraction of Common Contacts

We define our similarity measure, , as the fraction of common contacts between structures *A* and *B* with respect to the total number of contacts in *A*:

$$\mathrm{FCC}_{\mathrm{AB}} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}|} \in [0,1]$$

The outcome is a value ranging from zero, when the structures share no contacts, to a maximum of one when all contacts of structure *A* are present in structure *B*. The normalization of the number of common contacts over the number of contacts of the first structure brings asymmetry to the similarity measure and consequently to the similarity matrix since might not be equal to . In principle, the matrix could be symmetrized before clustering. However, a comparison of the clustering coverage and entropy of the obtained clusters using the two different matrices revealed that, for the majority of the cases, the symmetric matrix produces larger clusters but also with a larger entropy (Figure S4). In addition, the averaging of both FCC values reduces the resolution of the matrix, making it harder to optimize the clustering threshold. In light of these observations, the asymmetric matrix approach was chosen for all subsequent work.

In the case of symmetrical complexes, the chain identifier is omitted from the contact string identifier for the FCC calculation. This "chain-agnostic" variant of the FCC computation allows efficient clustering of structures that share the same interface regardless of the permutation of their chains along the symmetry axis.

Clustering Algorithm

We adapted a version of the disjoint Taylor-Butina clustering algorithm developed to use asymmetric matrices[205] that can be described in four steps:

1. Create a nearest-neighbor table from the full similarity matrix using a pre-defined threshold for the fraction of common contacts. Structure A is a neighbor of B only if is above the threshold, and vice-versa.

2. Detect true singletons (structures with an empty nearest-neighbor list, i.e., no neighbors at this threshold) and remove them from the dataset.
3. Find the structure with the largest nearest-neighbor list and define it as the center of the first cluster. Exclude this structure and all its neighbors from the dataset and update all nearest-neighbor lists. This update step is crucial to have disjointed clusters, or in other words, to ensure that structures belong to one and only one cluster.
4. Repeat step 3 until no structures are left in the dataset or the remaining have nearest-neighbor lists shorter than a pre-defined minimum cluster size threshold (default 4).

The original algorithm by Prinzie and Van der Poel[205] comprised the inclusion of the remaining structures – false singletons – in the cluster with the largest number of structures neighboring them. Preliminary analysis revealed that it contributed to an increase in the structural variability within the clusters, whilst being not at all justified with a significant increase of cluster population. Consequently, we forfeited this step in our implementation.

Implementation of the FCC-based Clustering Algorithm

Our clustering algorithm was implemented in the Python programming language and is freely available upon request. All the calculations were performed on a standard desktop computer with 2.66GHz CPU and 4GB RAM.

Measures for Cluster Quality Assessment

The quality of the cluster *i* can be assessed by the conformational variability of its N members. In the case of complexes, it is defined as the mean interface positional root mean square deviation (iRMSD) of all members from the center of the cluster, *clus.ctr*:

$$\text{Cluster}_i\ \text{Entropy} = \frac{1}{N} \sum_{s=1}^{N_i} \text{i-RMSD}(s)_{\text{clust.ctr}}$$

This measure was further expressed to account for the entropy of a given clustering run consisting of M clusters as the population-weighted average of the individual cluster entropies:

$$\text{Average Cluster Entropy} = \frac{\sum_{i=1}^{M} S(i) \times \text{Cluster}_i\ \text{Entropy}}{\sum_{i=1}^{M} S(i)}$$

where *S(i)* represents the number of elements in cluster *i*. Due to the internal symmetry of some cases of our structure set, calculating their cluster and average cluster entropies required an iterative iRMSD calculation where all chain combinations were tried. The lowest iRMSD value was then chosen. This ensured that, despite chain permutations, the entropies truly reflected the conformational variability within the clusters, while free from symmetry-induced artifacts.

Table 1. Biomolecular complexes used to assess FCC clustering performance. The models were taken from previously published datasets. The references of the experimental structure determination protocols are shown in parenthesis after the PDB ID.

Complex	PDB ID	No. of Components	Type	No. of Models	Symmetry
E2A/HPR	1GGR [206]	2	Prot./Prot.	200	none
Barnase-Barstar	1BRS [207]	2	Prot./Prot.	200	none
TBEV	1SVB [208]	3	Prot./Prot.	400	C3
LecB	1OUS [187]	4	Prot./Prot.	400	D2
VP1	1VPN [188]	5	Prot./Prot.	400	C5
PVUII/DNA	1EYU [209]	2*	Prot./DNA	200	none

* The focus in this complex was on the protein-DNA interface. Accordingly, we did not consider the protein-protein interface for clustering purposes.

Definition of iRMSD and iLRMSD

iRMSD: the interface RMSD is defined following CAPRI standards as the positional root mean square deviation of all interface residues (calculated on the Cα, N, C and O atoms) that have a heavy atom within 10Å of any other interacting partner.

*l*RMSD: the ligand RMSD is also defined following CAPRI standards. The models are first fit onto the larger chain (receptor) and then the RMSD (on Cα, N, C and O atoms) is calculated on the smaller chain (ligand).

ilRMSD: the interface-Ligand-RMSD (iLRMSD) used in HADDOCK for clustering purposes[59] is a slight variation of iRMSD, in which the models are first fit on the interface of the first molecule and the RMSD is then calculated on the interface residues of all other molecules. Interface residues are automatically defined based on all contacts observed over all generated docking solutions. For speed purposes only CA atoms are considered. Depending on the conformation sampling of the docking models, this measure will be somewhere in between interface- (sampling restricted to a particular region in the receptor surface) and ligandRMSD (sampling of the entire surface of the largest molecule).

Results

Evaluating FCC as a similarity descriptor

In order to evaluate the performance of FCC clustering, we analyzed docking models obtained with HADDOCK[59] for a set of six complexes consisting of two to five components and with various internal symmetries (see Table I and Table II) whose structures were experimentally determined. We calculated both the fraction of common contacts with the reference native structure (FCC_{NAT}) and the interface positional RMSD (i-$RMSD_{NAT}$) from the native structure (see Material and Methods). Additionally, we also calculated the ligand RMSD (L-$RMSD_{NAT}$) from the native structure.

Unsurprisingly, for all complexes, near-native models (low i-$RMSD_{NAT}$) share a substantial number of contacts (high FCC_{NAT}) with the native structure, while those more dissimilar share progressively fewer or none at all (Figure 1). The same, albeit less obvious for some structures, is observed for L-$RMSD_{NAT}$ (Figure S1). The anomaly observed for

the LecB protein, a dimer of dimers, is due to its particular symmetry type (D2), in which the larger intra-dimer interface accounts for the majority of contacts, causing solutions mirrored across the inter-dimeric axis to have high FCC values (Figure S2). Nevertheless, the most native-like structures have a distinctly higher FCC value. These observations indicate that FCC is a good similarity descriptor, suitable for clustering of biomolecular interfaces, regardless of their molecular components and their quaternary arrangement.

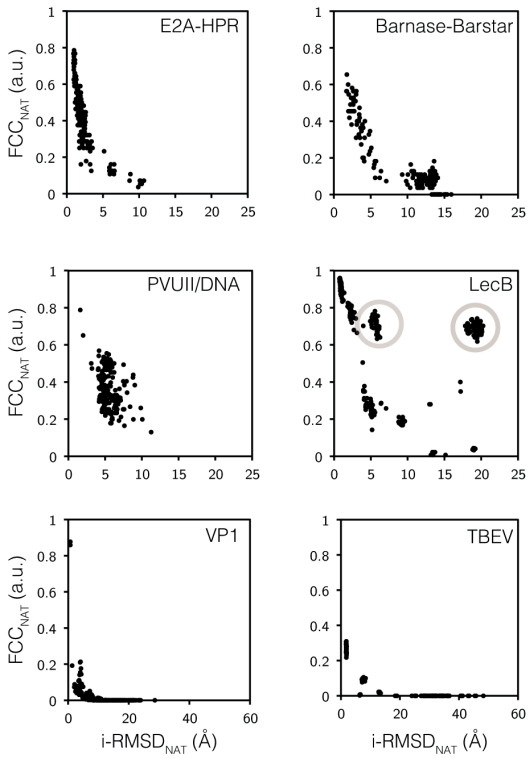


Figure 1. Assessing the fraction of common contacts (FCC_{NAT}) as a similarity descriptor by comparison with the interface positional root mean square deviation ($i\text{-}RMSD_{NAT}$). Both FCC and i-RMSD are calculated with respect to the experimentally determined structure of each complex. Low $i\text{-}RMSD_{NAT}$ values correspond to high FCC_{NAT} values, which supports the hypothesis that FCC is a good similarity descriptor and hence, a good similarity measure for structural clustering. In the case of LecB, symmetrical solutions that share only the larger of the two dimeric interfaces with the native structure have high FCC values (highlighted by gray circles) (see also Figure S2 and the main text for explanation).

Choice of optimal threshold for FCC-based clustering

The clustering threshold defines the rigor with which the clustering algorithm considers two structures similar enough to belong together in the same cluster. While previous works[199] have attempted to derive an optimal threshold for protein-protein docking using the distribution of values in the similarity matrix, pursuing a similar approach for FCC proved unreasonable. The distribution of the values in the similarity matrix depends on the conformational variability of the generated models. For complexes whose models are widespread over the conformational landscape, such as in the majority of our structure set (Table II), the distribution of similarity matrix values resembles a negative exponential function (Figure S3, panels B, C, D and E) and is therefore unsuitable for extracting an optimal clustering threshold as per *Kozakov et al.*[199]. Interestingly, the chain-agnostic variant of the algorithm affects the distribution of the matrices of multimeric complexes,

producing a shift towards higher FCC values (Figure S3, panels C, D and E). In light of these observations, we opted to evaluate several clustering runs at different thresholds (starting at 0.5, with increasing steps of 0.1) monitoring the conformational entropy of the resulting clusters and the total percentage of models included in clusters (Figure 2).

As expected, raising the clustering threshold enhances structure discrimination. This increasingly isolates structures and consequently reduces the size of the resulting clusters. Eventually, these clusters fail to meet the minimum size requirement (4 members) and their members are considered isolated events in the conformational landscape (Fig 2B). This effect is particularly evident at very highly discriminative thresholds (0.9) where the fraction of clustered structures drops below 0.5 for most cases.

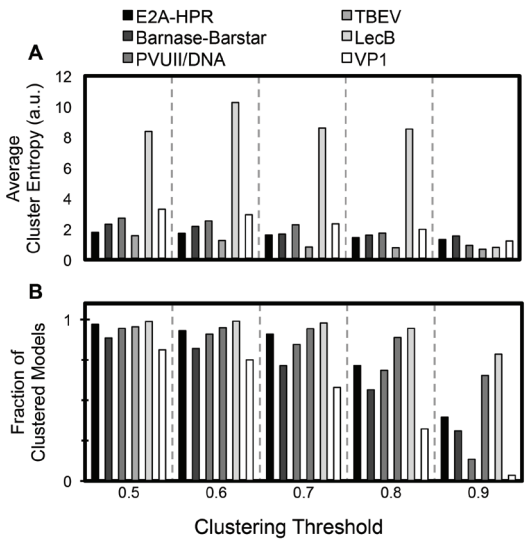


Figure 2. Definition of an optimal clustering threshold from an analysis of different runs at different clustering thresholds. A value of 0.75 was selected based on the observation that the entropy of the clusters declines with increasing values of threshold while the number of structures included in the clusters only drops sharply at 0.9. This threshold is appropriate for good clustering in all cases but LecB, which requires a higher value (0.9) due to its particular symmetric arrangement.

The average cluster entropy depends on the quality of the docking prediction and on the dispersion of the models over the conformational landscape of the molecule. Well-defined model sets such as E2A-HPR produce clusters through the FCC algorithm with an entropy comparable to those of RMSD clustering at thresholds as low as 0.5 (50% of the interface contacts in common) (Figure 2A). Stricter discrimination has little effect on the structural variability in each cluster, as seen by the slow decrease in average cluster entropy of E2A-HPR (1.78 to 1.32). On the other hand, clustering more chaotically distributed model sets (e.g. VP1) clearly benefits from higher thresholds, since the entropy of the resulting clusters steadily drops (from 3.30 to 1.22) as the threshold increases. Notably, LecB deviates from the rest of the complexes due to its particular symmetrical arrangement (Figure S2). Up to a threshold of 0.8, most clusters include several mirror-like symmetrical conformations and have consequently very high entropy values (>10). Increasing the threshold to 0.9 allows the discrimination of both inter-dimer and intra-dimer interfaces, splitting the very large cluster obtained at 0.8 (Cluster #1, entropy 10.55, N=297) into smaller but extremely compact sub-clusters. This brings the average cluster entropy sharply down (0.91) while

retaining the large majority of the structures (78.5%) (Figure 2). These observations suggest that a threshold between 0.7 and 0.8 – empirically, 0.75 - is the most suitable for generic application of FCC clustering. This might however require adaptation in particular cases, such as LecB.

Quantitative assessment of FCC clustering

To cement the quality of FCC as a valid similarity measure for structural clustering, we performed a direct comparison with the protocol integrated in HADDOCK, which uses interface-ligand root mean square deviation of atomic coordinates (iL RMSD, see Material and Methods) and the clustering algorithm implemented by Daura[210] with a default clustering threshold of 7.5Å (Figure 3). iL RMSD clustering at this threshold collects a larger number of structures at an expected cost of higher entropy clusters (Figure 3A and 3B). While for the heterodimers and PVUII/DNA this is acceptable, analysis of clustering of symmetric multi-component complexes reveals an important limitation of iL RMSD, and by extension all positional RMSD based metrics, as clustering similarity measures: recognizing similar conformations with different symmetrical chain arrangements is not trivially possible and results in several clusters that should, in truth, be merged. This happens since these methods are bound to the chain identifiers of the PDB file format, which in turn results in high RMSD values for structures that share very similar structural features but whose chain identifiers are swapped, placing them in separate clusters. Detailed analysis of the centers of iL RMSD-generated clusters corroborates this hypothesis, showing little conformation differences between several models, indicating that these should belong in the same cluster. By contrast, the chain-agnostic variant of the FCC clustering algorithm agglomerates the several chain permutations (i.e. for a three-chain complex: ABC, ACB) in one single and larger cluster (N=87). Since these structures are nevertheless very similar, the entropy of the clusters remains extremely low (Figure 3B). Finally, since structural alignment and fitting is absent in FCC clustering, computational efficiency is greatly enhanced (Figure 3C): iL-RMSD clustering takes several minutes to several hours to build similarity matrices for the complexes, depending on the interface size. Using FCC as a similarity measure reduces this computation time by a factor of, on average, 100, smoothing the path for structural clustering of intricate multi-component systems such as those described before.

Discussion

Residue contacts are enough to differentiate binding poses

We have developed a new clustering approach for macromolecular complexes based on the premise that residue contacts alone are enough to discriminate binding poses between interacting partners. Although already used by the docking community to assess the accuracy of the docking results, the application of the fraction of common contacts (FCC) in structural clustering of docking solutions is novel and shows good results. Direct comparison with the commonly used interface RMSD (i-RMSD) reveals that FCC is a good descriptor of structural similarity (Figure 1). We have shown that a high value of FCC unequivocally corresponds to low i-RMSD and L-RMSD values, and therefore similar structures, independently of the number of components in the complex or its symmetry type.

Table 2. Structural characteristics of the model set and clustering statistics for both RMSD and FCC clustering methods. The mean interface RMSD of the ensemble informs on the variability of the conformations present in the model set. The percentage of clustered structures, and average cluster entropy, refer to two measures we defined to assess the clustering algorithms. FCC clustering is shown to have consistently lower entropy, at a cost of less structures clustered, and performing particularly well for multimeric assemblies (TBEV, LecB, and VP1).

Complex	Mean i-RMSD of the ensemble (Å)	Similarity Measure	No. of Clusters	Clustered Structures (%)	Average Cluster Entropy (a.u.)
E2A-HPR	2.4 ± 1.8	RMSD	3	99	1.6
		FCC	7	83	1.7
Barnase-Barstar	8.9 ± 4.0	RMSD	8	92	1.9
		FCC	7	65	1.8
TBEV	27.2 ± 11	RMSD	22	94	1.7
		FCC	17	92	0.9
LecB	8.9 ± 7.7	RMSD	25	93	1.9
		FCC	18	79	0.8
VP1	15.0 ± 5.3	RMSD	32	72	2.7
		FCC	32	50	2.1
PVUII/DNA	5.57 ± 1.3	RMSD	8	93	2.2
		FCC	11	80	2.1

FCC clustering can accommodate various levels of biomolecular complexity

We have shown that FCC clustering deals effortlessly with large assemblies, greatly reducing the computation time while generating clusters of similar quality with the current state-of-the-art methods (Figure 3). This leap in performance is due to the avoidance of pair-wise structural alignments, which has the added value of removing the bias stemming from the choice of regions on which to perform the alignment. Another problem tied to structural alignment lies in the handling of symmetrical solutions. While structural biologists artificially name molecular chains to distinguish them from one another, for structural comparison and by proxy, structural clustering purposes, the chain arrangement does not matter as long as the molecular architecture is similar. Since RMSD calculations are bound to the chain identifiers, clustering based on such measures often produces very similar clusters whose structures differ only in the symmetrical arrangement of their chains. This is evident in all the cases with internal symmetry presented above (TBEV, LecB, VP1) and poses a problem for post-clustering analysis. Avoiding this problem in RMSD-based methods requires an iterative calculation of all the several permutations of the chain arrangements (e.g. ABC, ACB), which further aggravates computational performance. FCC clustering sidesteps all these issues by considering each complex a whole entity free from chain identifiers – the chain-agnostic variant. Although simplistic, this solution successfully merges the several clusters that share the same conformation, which not only accounts for larger clusters but also facilitates posterior analysis, namely in determining the lowest energy cluster, likely to contain the best representative structure. This advantage will be crucial in case where only few similar conformations are present in the model set. Furthermore, since the calculation of the fraction of common contacts, as per the current algorithm,

reads only the residue index within the structure, FCC has a wide range of applications regarding different molecular representation scales (coarse-grained to all-atom). It also allows for clustering of point mutants of the same structure, or even gapped models, given that the numbering is preserved and consistent across all models.

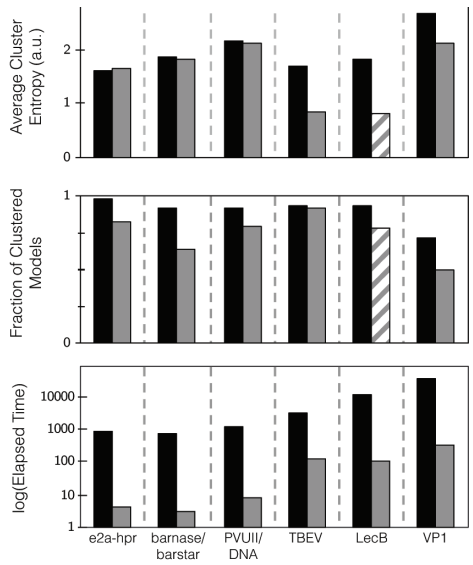


Figure 3. Average cluster entropy, cluster coverage, and computational performance for both the fraction of common contacts clustering (FCC) (gray bars) and the interface-ligand RMSD (iL RMSD) clustering (black bars). Clusters were generated using the default threshold for iL RMSD of 7.5Å in all cases, and 0.75 for FCC, except LecB, which was clustered at 0.9 (striped bar). FCC clustering leads to smaller but more compact clusters, which indicates a better discrimination of fringe structures. For multimeric structures with internal symmetry, in particular TBEV and VP1, the advantage of clustering based on FCC is evident. Performance-wise, avoiding structural fitting reduces the computation time required for FCC clustering by a factor 100 on average.

Ranking of clusters is largely independent of the clustering method

The discriminative power of FCC clustering for the chosen general threshold of 0.75 (75% of the interface in common) is superior to that of iL-RMSD clustering, reducing the entropy of the clusters, but also the size of the clusters. Analysis of which structures are effectively discarded through FCC clustering showed that these are largely fringe structures, the furthest away from the cluster center, and that in most cases do not impact the overall quality of the clusters when compared to the native structure. An analysis on the average i-RMSD$_{NAT}$ of all clusters generated with both FCC and iL-RMSD algorithms for an extended dataset composed of twenty real-case scenarios (previous CAPRI experiment targets) showed that the ranking of the clusters is largely unaffected by the clustering method (Table S1). Comparison of the the top ranking clusters reveals in a majority of cases a good agreement between both clustering algorithms and for a number of cases, for similar ranking performance, the resulting clusters show an increased accuracy as measured by i-RMSD$_{NAT}$. Therefore, this corroborates that FCC clustering is not discarding important native-like structures and is therefore suitable for large-scale application.

FCC clustering accommodates current and future needs in biomolecular docking

Both the increased computational efficiency and the overall performance of our FCC clustering algorithm are encouraging. Efficient methods that allow for rapid RMSD calculation of protein complexes exist but are, however, mostly based on simple rigid body transformations (i.e. rotations and translations over the center of mass of the complex) and

thus do not account for internal flexibility of the system. Since the models were generated with HADDOCK, which includes a semi-flexible refinement step, rigid body-based clustering algorithms are inappropriate. In contrast, FCC clustering worked effectively on these models, meaning that the method is suited for flexible docking approaches, without degrading performance. Furthermore, considering the shift toward the modelling of entire interactomes or very large systems (e.g. nuclear pore[197]) to fill in the gaps left by low-resolution or high-throughput experimental techniques, fast and accurate clustering methods will be critical in the near future. We have demonstrated here that our FCC algorithm is well-suited for this task as it performs well in diverse environments, from traditional protein-protein complexes to more complicated multi-component assemblies and heterogeneous biomolecular systems like protein-DNA complexes, while being computationally efficient.

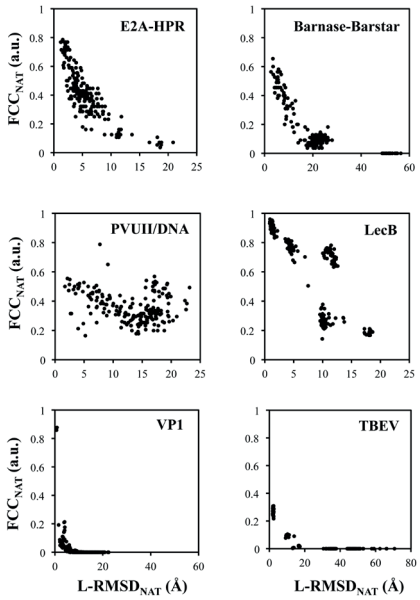
Conclusion

The current perspectives for the field of biomolecular docking call for methods able to deal with large datasets, both in number of molecules and molecular size. RMSD-based clustering methods are computationally expensive and their sensitivity decreases with the molecular size of the system. Yet, suggested alternatives so far, although useful in particular scenarios, fail at reproducing both their quality and performance when applied generically. While the concept of contact-based molecular comparison is known and used in both CASP and CAPRI, it is limited to the assessment of results. The inclusion of FCC clustering in docking algorithms, as shown here with HADDOCK, has the potential to greatly enhance their computational performance. In addition, FCC clustering is able to deal with symmetry and multi-component complexes with negligible performance degradation. Furthermore, given its sole dependence on residue numbering, it allows for the clustering of mutants and gapped structures, broadening even more its usefulness to the clustering of structures coming from different trajectories or simulations. All these, allied to the simplicity of the algorithm and its flexibility in dealing with several molecule types, tailor FCC clustering for the upcoming challenges in the docking field and offer an effective alternative to traditional RMSD-based clustering methods and their inherent shortcomings.

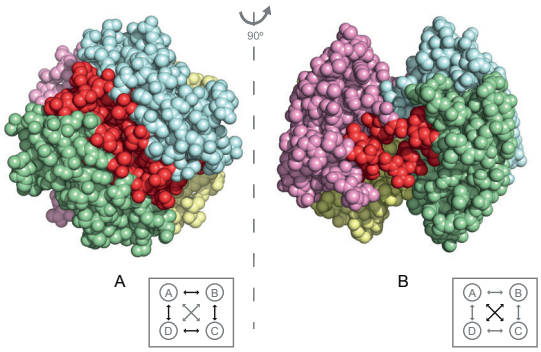
Acknowledgements

The authors thank M. van Dijk (University Utrecht) for helpful discussions.

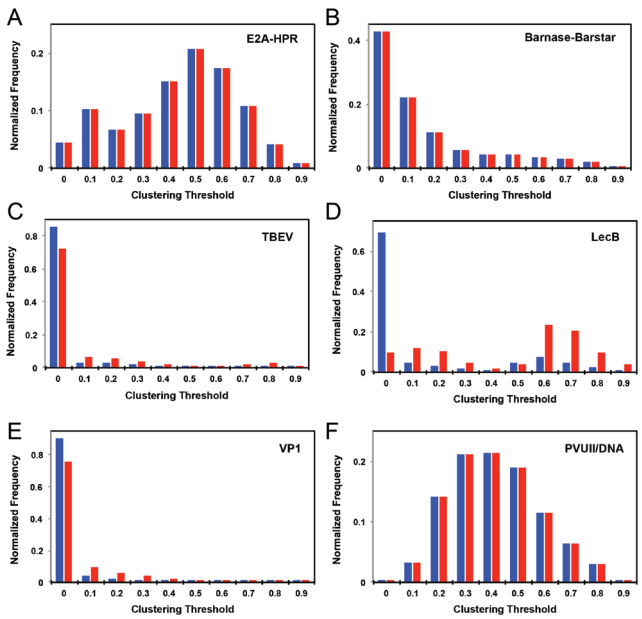
Supplementary Figures



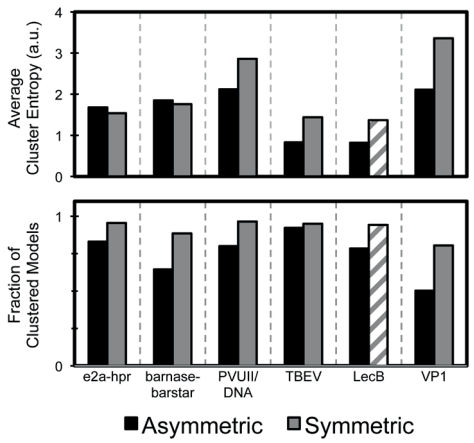
Supplementary Figure 1. Assessment of the discriminative capacity of FCC. The L-RMSD is calculated with respect to the experimentally determined structure of each complex (see Material and Methods). As with i-RMSD$_{NAT}$, low L-RMSD$_{NAT}$ values correspond to high FCC$_{NAT}$ values, which further corroborates the descriptor quality of FCC.



Supplementary Figure 2. Interfaces of LecB (PDB ID: 1OUS). LecB is composed of four monomers, arranged in two dimers with D2 symmetry. There are two easily distinguishable interfaces in the molecule, highlighted in red: **(A)** the larger intra-dimer interface (AB, AD, BC, DC) accounts for 532 atomic contacts (calculated at a threshold of 5Å), while **(B)** the smaller inter-dimer interface (AC, DB) accounts for only 105 contacts. The thick arrows in the boxes indicate the respective interfaces. This discrepancy in the number of contacts between the two interfaces, together with the symmetric restraints used to build the model in HADDOCK, results in solutions that are opposite mirror-images (i.e. ABCD and CDAB); since their intra-dimer interfaces are conserved, these have a high fraction of contacts in common, while their calculated interface RMSD is approximately ~20Å. This artifact explains the sudden drop in cluster entropy at threshold 0.9; the smaller inter-dimer interface is then discriminated.



Supplementary Figure 3. Histograms of the Similarity Matrix elements obtained with FCC and both algorithm variants: canonical in blue and chain-agnostic in red. Structures whose models are scattered all over the conformational landscape have a decreasing profile with a maximum at FCC=0 **(B, C, E, D)**. Narrowly sampled structures **(A, F)** on the other hand have a more harmonized distribution of FCC values. While for non-symmetrical complexes **(A, B, F)** the distribution is the same independently of the variant of the algorithm, for symmetrical complexes there is a shift of the matrix distribution towards higher values in the chain-agnostic variant, reflecting a better identification of similar overall conformations, irrespective of the chain arrangement.



Supplementary Figure 4. Comparison of the results of FCC clustering with an asymmetric similarity matrix (threshold 0.75, 0.9 for LecB) and a symmetric matrix (threshold 0.5, 0.8 for LecB). The symmetric matrix was derived from averaging the FCC values of the regular asymmetric matrix. For all cases, the symmetric matrix allows for more structures clustered. However, in most cases, this causes also an undesirable increase in entropy. Fine-tuning the clustering threshold was also problematic due to the discretization of the values in the similarity matrix caused by averaging. This effect makes small changes in the threshold to influence drastically the results, such as a sharp drop in cluster coverage at increasing thresholds, to well 50% in some cases at 0.75 for example.

Supplementary Tables

Supplementary Table 1. Assessment of the influence of the clustering algorithm in the ranking of final water refined structures for an extended benchmark of 'real-case' CAPRI targets. We generated clusters using a pre-defined generic threshold for FCC (0.75, see main text) and the default threshold in HADDOCK for iL-RMSD (7.5Å). The clusters are ranked according to their HADDOCK score, to mimic what a potential user would receive as output from the software. Only clusters with an interface RMSD lower than 4Å were kept (acceptable as per CAPRI criteria). Targets 19, 24, 30, 32, 38, and 39 failed to produce any clusters with acceptable solutions for both algorithms. There are only little differences between both methods, demonstrating that FCC clustering does not considerably changes the ranking of the final clusters (or even seems to improve the quality of the generated clusters) and is therefore suitable for usage in generic scenarios.

Complex	iL-RMSD Clustering (7.5Å)			FCC Clustering (0.75)		
	Rank	i-RMSD$_{NAT}$	FNAT	Rank	i-RMSD$_{NAT}$	FNAT
CAPRI 10	1	1.76±0.03	0.26±0.02	1	1.75±0.04	0.28±0.01
	2	1.76±0.06	0.28±0.01			
CAPRI 11	-	-	-	2	3.22±0.19	0.49±0.03
	-	-	-	9	3.79±0.16	0.11±0.02
CAPRI 12	1	0.84±0.16	0.82±0.05	1	0.84±0.16	0.82±0.05
	6	2.12±0.59	0.57±0.22	5	3.92±0.24	0.26±0.03
	-	-	-	6	1.88±0.4	0.58±0.12
CAPRI13	1	3.85±0.75	0.29±0.09	1	3.11±0.19	0.23±0.02
	2	1.73±0.69	0.52±0.11	2	1.35±0.19	0.58±0.03
	3	3.25±0.09	0.23±0.02	4	3.08±0.13	0.34±0.06
	-	-	-	7	4.32±0.50	0.18±0.03
CAPRI 14	1	1.08±0.1	0.56±0.01	1	1.08±0.10	0.56±0.01
	3	3.78±0.65	0.32±0.02	3	3.49±0.25	0.31±0.03
CAPRI 15	1	3.88±0.93	0.32±0.05	2	4.86±0.92	0.22±0.08
	7	3.76±0.67	0.25±0.03			
CAPRI 18	1	1.99±0.47	0.67±0.12	1	1.99±0.47	0.67±0.12
	6	3.14±0.33	0.30±0.01	5	3.14±0.33	0.30±0.01
CAPRI 19	-	-	-	-	-	-
CAPRI 21	1	2.11±0.15	0.46±0.09	1	2.16±0.18	0.51±0.10
	7	3.44±0.17	0.34±0.04	4	2.93±0.35	0.38±0.05
	9	3.05±0.73	0.36±0.12	5	2.96±0.33	0.39±0.08
	-	-	-	8	3.13±0.54	0.29±0.11
CAPRI 24	-	-	-	-	-	-
CAPRI 25	2	3.24±1.25	0.43±0.17	2	2.16±0.61	0.62±0.11

Supplementary Table 1 (continued). Assessment of the influence of the clustering algorithm in the ranking of final water refined structures for an extended benchmark of 'real-case' CAPRI targets.

Complex	iL-RMSD Clustering (7.5Å)			FCC Clustering (0.75)		
	Rank	i-RMSD$_{NAT}$	FNAT	Rank	i-RMSD$_{NAT}$	FNAT
CAPRI 26	2	2.33±0.5	0.52±0.07	1	2.48±0.35	0.49±0.03
	-	-	-	3	1.63±0.26	0.55±0.05
	-	-	-	4	2.66±0.24	0.41±0.04
	-	-	-	6	2.16±0.27	0.47±0.05
	-	-	-	10	2.06±0.15	0.44±0.04
	-	-	-	11	2.22±0.26	0.42±0.08
	-	-	-	19	2.57±0.21	0.34±0.03
CAPRI 27	1	2.28±0.34	0.60±0.05	1	2.28±0.34	0.60±0.05
	4	4.08±0.60	0.43±0.09	3	3.56±0.29	0.27±0.08
	-	-	-	4	3.71±0.18	0.43±0.07
	-	-	-	7	3.76±0.49	0.32±0.03
CAPRI 29	1	2.05±0.51	0.51±0.11	2	2.44±0.29	0.46±0.10
	4	3.19±0.40	0.31±0.02	3	1.77±0.24	0.60±0.03
	-	-	-	6	1.77±0.24	0.46±0.04
	-	-	-	7	2.49±0.32	0.36±0.03
	-	-	-	8	1.90±0.05	0.50±0.06
	-	-	-	10	1.83±0.19	0.40±0.04
	-	-	-	11	3.70±0.24	0.28±0.03
CAPRI 30	-	-	-	-	-	-
CAPRI 32	-	-	-	-	-	-
CAPRI 37	2	2.18±0.89	0.65±0.16	2	2.92±0.43	0.52±0.06
	6	3.72±0.84	0.41±0.07	-	-	-
	9	2.40±0.85	0.57±0.14	-	-	-
CAPRI 38	-	-	-	-	-	-
CAPRI 39	-	-	-	-	-	-
CAPRI 40A	1	1.07±0.06	0.86±0.01	1	1.07±0.06	0.86±0.01
CAPRI 40B	1	1.01±0.11	0.80±0.04	1	1.01±0.11	0.80±0.04
CAPRI 41	2	2.77±0.15	0.48±0.04	3	2.79±0.13	0.50±0.04
	-	-	-	7	2.45±0.18	0.42±0.06
CAPRI 42	6	1.75±0.56	0.71±0.25	6	1.40±0.13	0.86±0.03
	7	3.82±0.39	0.22±0.06	7	4.23±0.55	0.25±0.04
	-	-	-	8	3.46±0.19	0.17±0.03

Supplementary Table 1 (continued). Assessment of the influence of the clustering algorithm in the ranking of final water refined structures for an extended benchmark of 'real-case' CAPRI targets.

Complex	iL-RMSD Clustering (7.5Å)			FCC Clustering (0.75)		
	Rank	i-RMSD$_{NAT}$	FNAT	Rank	i-RMSD$_{NAT}$	FNAT
CAPRI 50	1	2.46±0.79	0.46±0.10	1	3.30±0.23	0.36±0.05
	3	2.76±0.28	0.41±0.06	2	2.76±0.28	0.41±0.06
	6	3.99±0.38	0.13±0.10	4	1.47±0.11	0.70±0.02
	-	-	-	6	1.81±0.54	0.61±0.08
	-	-	-	7	1.37±0.54	0.63±0.11
	-	-	-	8	3.27±0.37	0.27±0.03
	-	-	-	11	2.53±0.35	0.32±0.08
	-	-	-	12	3.77±0.24	0.18±0.02
	-	-	-	13	3.89±0.38	0.09±0.02
	-	-	-	14	3.73±0.27	0.08±0.02
	-	-	-	16	2.83±0.50	0.39±0.13

Chapter 5

HADDOCK 3.0: Integrative modelling of supramolecular assemblies

J. Rodrigues*, E. Karaca*, C. Don*, HADDOCK Team, A.M.J.J Bonvin

** These authors contributed equally to this work.*

Unpublished

Abstract

For the last two decades, biologists have been challenging the limits of structure determination methods with larger and more complex biological systems. Molecular machines such as the ribosome have had their atomic structure resolved directly from x-ray diffraction, while others such as the nuclear pore complex were modeled based on a combination of experimental methods. In both examples, and many others, computational modeling played an important role in interpreting and translating the experimental data into three-dimensional models. It is therefore crucial that computational methods adapt to 1) combine different types of structural data and 2) support and efficiently process large and heterogeneous molecular systems. In this work, we present the new version of the HADDOCK integrative modeling software: HADDOCK 3.0. We rewrote the core of the software to provide a better user experience through validation routines, and simplify the integration of new features. More importantly, we implemented a new initial orientation protocol that allows simultaneous modeling of an arbitrary number of molecules and integrated a coarse-grained representation of proteins, through the MARTINI force field, to improve computational efficiency. These new features, combined with the possibility of using experimental and prediction data from numerous different sources, should benefit the structural biology community and consolidate the reputation of HADDOCK as a state-of-the-art modeling software for biological assemblies.

Introduction

In the molecular biology of the cell, the whole is greater than the sum of the parts. Each individual molecule, whether a protein or a nucleic acid, has a particular function, but it is in their association into complexes that lies the driving force being important processes such as cell signaling and proliferation[109]. Consequently, dissecting these molecular assemblies and their interactions, particularly at the atomic-level, is paramount for the understanding of our own biology, including that of diseases[1,211].

The character of the interaction depends on its function[18]: ubiquitin signaling, for example, is driven by several highly specific, somewhat weak (dissociation constant in the micromolar range) and short-lived interactions[103]. On the other hand, (most of) the 21 protein subunits that constitute the core of the 30S subunit of the eukaryotic ribosome assemble efficiently and rather quickly *in vitro*. Since either type of interactions presents, respectively, a specific challenge to the golden duo of structure determination, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, other complementary experimental methods, such as cryo-electron microscopy and small-angle scattering, have been frequently employed in structural studies of biological assemblies. Cryo-electron microscopy, in particular, although traditionally low-resolution, has shown the potential to provide atomic-level structures of extremely large assemblies[212].

This collection of techniques has made the last four decades of structural biology remarkably productive[213]: more than one hundred thousand structures have been deposited into the Protein Data Base (PDB). The size and complexity of these structures is also growing with each passing year, reflecting the continuous development and improvement of molecular and structural biology techniques. Although the majority of the depositions is still of monomeric proteins, large molecular machines such as the ribosome[212,214] and the spliceosome[215] are being characterized at an unprecedented level of detail.

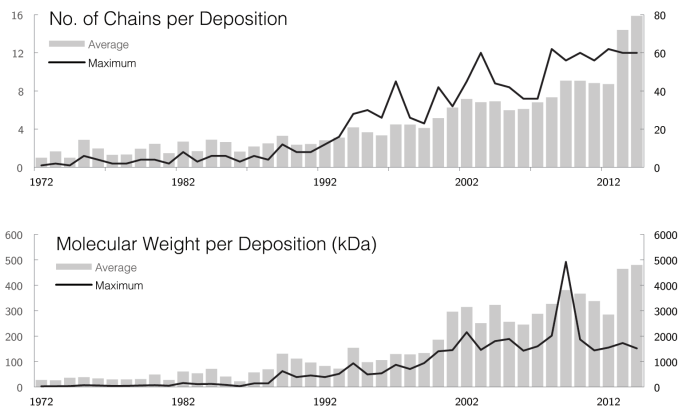


Figure 1. Statistics from 101208 structure depositions available at the RCSB PDB website (accessed on the 1st of July 2014). The trend for structure determination of larger and more complex systems, both by the improvement of high-resolution x-ray crystallography and NMR, as well as by the development of low resolution structure determination methods such as cryo-electron microscopy and small-angle scattering over the last twenty years is quite clear.

This rapid explosion of the structural universe is partially owed to developments in the field of computational structural biology, whose methods allowed the translation of data from the different structure determination techniques into chemically sound three-dimensional models[216]. Computational methods also enable the prediction of the structure of biological systems for which there is no experimental structure available[13,217], which is often extremely helpful in formulating hypothesis and interpreting results. More interestingly, in the last years, these methods allowed researchers to combine data from individual experiments and build reasonably accurate models of biological systems that would otherwise remain unknown (e.g. the nuclear pore complex[197,218]). This class of integrative or data-driven computational methods, has recently been experiencing a surge in popularity, being hailed as the key to unlock structural and functional information of many hitherto unreachable biological systems[14,219].

HADDOCK (High Ambiguity Driven DOCKing), one of the pioneers in integrative methods for modeling protein interactions (docking), was originally designed to handle highly ambiguous NMR and mutagenesis data[30]. At its core is the calculation engine CNS (Crystallography and NMR System) [64,128], that supports a variety of NMR and x-ray energy terms and allows the introduction of explicit flexibility during refinement, enabling the modeling of conformational changes[16,34]. Another powerful feature of HADDOCK is the concept of ambiguous interaction restraints (AIRs), where surface patches that are potentially involved in the interaction between the molecules are included in the modeling process as highly ambiguous distance restraints[220]. This powerful approach fits perfectly the ambiguous nature of chemical shift perturbation data from NMR, often used to probe the interface of protein complexes, as well as noisy data from bioinformatics predictions[58]. Over the years, support was added for orientational restraints[65,66,221] and shape information from SAXS and ion mobility mass spectrometry[67]. HADDOCK also supports data from any experimental method that provides distance information: e.g. crosslinks detected by mass spectrometry, Förster resonance energy transfer (FRET) and electron paramagnetic resonance (EPR). Parallel to the integration of new types and sources of data, other developments have focused on improving the modeling of other biological systems, such as protein-DNA [222,223], protein-peptide[70], and multi-body complexes[69]. In addition, the description of biological interactions was also improved by the development of solvated docking protocols that explicitly account for interfacial water molecules[141,142]. Altogether, the combination of data, powerful algorithms, and an easy-to-use web interface[118] has positioned HADDOCK as particularly attractive solution for the scientific community at large[57].

Despite its success, the current implementation of HADDOCK suffers from a few shortcomings that will become increasingly apparent with the current trend of larger and more complex structural studies. As of now, it is possible to model in HADDOCK the interaction of up to six individual molecules, although this limit can be somewhat overcome by merging several protein chains into one, as done recently in a structural study of the circadian clock[224]. The number of different restraint sets that researchers can use in the calculations is also hard-coded to some extent. Another limitation is the computational cost of simulating large molecular machines, which might require several weeks of calculations on a moderate-sized cluster. Finally, the current implementation of the software demands some technical expertise in order to be properly used, namely in setting up a modeling calculation for non-standard systems not supported by the web server implementation. As such, the application of HADDOCK to very large molecular machines is hindered, if not outright impossible.

This chapter describes the complete rewrite of HADDOCK, which we name version 3.0. Novel features include the implementation of an initial orientation algorithm to distribute molecules in space prior to docking and the generalization of the entire codebase for an arbitrary number of molecules and restraint sets. To tackle the molecular weight / system size limitation, we implemented the MARTINI coarse-grained force field in CNS and an efficient all-atom conversion step that accounts for potential conformational changes in the coarse-grained models at the end of the calculations. Lastly, we refactored the workflow of the software to include a validation step for the input data, until now present only in the web server interface, in order to flatten the learning curve for new users.

Material And Methods

Workflow of HADDOCK

A typical HADDOCK calculation starts by separating the molecules in space and randomly translating and rotating them within a 10Å-radius cube to remove any initial orientation bias and generate a starting model for the docking. This model undergoes several trials of rigid-body energy minimization using the Powell minimizer, each of which includes an additional minimization of a 180-degree rotated model around an axis perpendicular to the interface, followed by a scoring step. After every trial, HADDOCK separates and randomizes the molecules again. Each rigid-body model written to disk is the best scoring solution of the trials (by default ten - 5x2 rotated solutions). This procedure is repeated, with a different starting seed, as many times as required to produce the desired number of initial rigid-body models of the complex, by default 1000.

The second step of the modeling calculation applies only to the best scoring fraction of the rigid-body models, typically 200, and consists of a thorough semi-flexible refinement in torsion angle space of the interface region. Using a simple distance cutoff of 10Å between the interacting partners, HADDOCK identifies the interface atoms of each model and defines them as flexible while keeping all others frozen. Each model is subjected to three consecutive simulated annealing refinements: the first optimizes the orientation of the partners as rigid-bodies, the second refines the side-chains of the interface region, and the third allows for larger conformational rearrangements as both side-chains and backbone of the interface are refined. This protocol can typically target conformational changes of up to 2Å backbone interface RMSD[59].

The final refinement stage of HADDOCK immerses the models in an 8Å shell of explicit solvent (TIP3 water, as default, or DMSO) and runs molecular dynamics with weak position restraints, first on all non-interface heavy atoms and then only on non-interface backbone atoms, to further optimize the interface. Typically, there is no selection step between the simulated annealing refinement and the explicit solvent refinement, i.e., all models undergo both.

To facilitate the analysis, HADDOCK clusters all models and produces a final ranking based on the average score of the best four models of each cluster. For further details on this workflow, refer to the original publication of HADDOCK[30]. The three steps, rigid-body energy minimization, semi-flexible simulated annealing, and explicit solvent refinement are named *it0*, *it1*, and *itw*, respectively, for simplicity.

Throughout the modeling and at the final stage, HADDOCK scores the models based on a linear combination of several physics-based and knowledge-based energy terms. These comprise Coulomb electrostatics and van der Waals interactions, as parameterized in the

OPLS force field[151]; an empirical desolvation term[129]; the buried surface area of the complex; and the various restraint energies as implemented in CNS (e.g. AIRs, symmetry, NMR dipolar couplings). The different stages of the modelling attribute different weights to each of these terms, reflecting particular chemical properties of the association process (i.e. the initial attraction of the proteins is based on long-range electrostatics, which is replaced by fine-tuning of van der Waals interactions as the interfaces come together):

$$E(it0) = 0.1 \times E_{elec} + 0.01 \times E_{vdW} + 1.0 \times E_{desolv} + 0.01 \times E_{AIR} - 0.01 \times BSA$$

$$E(it1) = 1.0 \times E_{elec} + 1.0 \times E_{vdW} + 1.0 \times E_{desolv} + 0.1 \times E_{AIR} - 0.01 \times BSA$$

$$E(it0) = 0.2 \times E_{elec} + 1.0 \times E_{vdW} + 1.0 \times E_{desolv} + 0.1 \times E_{AIR}$$

Initial placement of molecules before docking

Before the first rigid-body energy minimization, HADDOCK separates the interacting molecules in space. This separation has two goals: first, to remove any orientation bias from the coordinates that the user submitted, and second, to place the molecules as equidistant from each other as possible, as to remove any starting position bias for the minimization. However, since there is an upper limit of n+1 equidistance points for an n dimensional system, we cannot satisfy this condition for systems of five or more molecules. Originally, HADDOCK performed this separation based on simple geometric rules that depended on the number of molecules: placing two molecules along a line; three along the vertices of an equilateral triangle; four on the vertices of a tetrahedron; six on the vertices of an octahedron. For large numbers of molecules, this solution depends on the availability of regular polyhedrons, which is obviously not optimal (5 molecules are placed on an octahedron minus one vertex, etc.) and is not easily generalizable for n molecules.

We based our general placement algorithm on the 'Thomson problem', or the uniform distribution of points on the surface of a sphere[225]. Given any number of points (N), we model them as particles of unitary negative charge (i.e. electrons) on a sphere, using the Golden Spiral algorithm, and minimize the Coulomb energy of the system using the steepest descent algorithm. In a few short steps, the points distribute themselves on the surface of the sphere to minimize electrostatic repulsion as best as possible. It is important to note that for N>50, there are multiple minimum energy configurations[226]. This is a well-known problem in mathematics, which we do not pretend to solve. However, for our purposes, the seemingly uniform distribution is a perfectly reasonable solution. The radius of the sphere can be scaled so that the minimum distance between any two points is at least D, which we define *a priori*. With this approach, it is possible to generate initial positions for an arbitrary number of molecules, while ensuring that their distribution is as unbiased as possible. The minimum distance between the molecules is set dynamically, based on the maximum dimension of the largest molecule plus a constant of 25Å (defined as roughly twice the cutoff for non-bonded interactions as defined in the force field). This ensures that for any molecular system, the docking of the molecules starts from unbiased positions. We refer to this approach as 'sphere-based placement' to distinguish it from the original approach.

Coarse-grained modeling with the MARTINI force field

In order to be of practical use, HADDOCK must run in a reasonable amount of time on modest computational resources (i.e. it is not realistic to expect HADDOCK to run quickly

on a laptop). For large systems, in particular, given the number of atoms involved, the calculations will be necessarily longer. A possible strategy is to reduce the resolution of the system, coalescing atoms into larger pseudo-particles. This approach, 'coarse-graining', was actually introduced in the early days of computational structure biology due to the limited computing power at that time[227]. Since then, it has been widely used to achieve larger time scales in molecular dynamics simulations[228,229] and also in protein docking to reduce the complexity of the conformational search space[32].

We chose to implement the well-accepted MARTINI (version 2.2) force field for proteins[230-232] in HADDOCK. MARTINI uses a four-to-one mapping: it replaces the four amino acid backbone atoms by a pseudo-atom bead and each side-chain by a maximum of to four beads. The position of each bead is the geometric center of mass of the atoms it is replacing. The parameterization of the side-chain beads is residue-dependent, while the backbone bead reflects the secondary structure to which a residue belongs.

The implementation of MARTINI in HADDOCK is as follows: first the secondary structure of the protein is calculated using DSSP[233] and the secondary structure elements converted to a numerical code (i.e. helix, 1; sheet, 2; …) which is encoded in the b-factor column of each amino acid residue in the PDB file. These values are used during the topology generation for each individual molecule to modify the parameters of the backbone bead accordingly. This avoids the otherwise very large number of parameters necessary to describe each amino acid in each secondary structure. By default, the coil parameters apply. We then convert the all-atom structure to a coarse-grained representation and produce a mapping of the atoms to the corresponding beads in the form of distance restraints. All MARTINI parameters were left as default, except for the conversion of the force constants from $kJ.mol^{-1}.nm^{-2}$ to $kCal.mol^{-1}.Å^{-2}$ and the occasional duplication of bead parameters to ensure their uniqueness as required by CNS.

There are no differences in the modeling workflow if the molecules are coarse-grained. However, since our aim is to provide atomic resolution models, we implemented a morphing protocol before the explicit solvent refinement. The coordinates of the all-atom molecules are first rigid-body minimized with respect to the distance restraints that map the atoms to the coarse-grained beads. Then, the all-atom models are 'morphed' onto the coarse-grained beads through a Powell energy minimization during which bonded, non-bonded, and distance restraints energy terms are active, followed by a short molecular dynamics simulation (500 steps). The conversion ends with the refinement of both molecules simultaneously, to both optimize the interface and remove potential clashes, using the same minimization and dynamics combination. The option of following through with the default protocol and performing the explicit solvent refinement is left to the user (by default on), as this might take a considerable amount of time, particularly in the case of very large systems.

Finally, since the desolvation energy term included in the HADDOCK score is dependent on the atom type and its surface area, it had to be adapted to also include the coarse-grained beads. For each residue, we calculated the desolvation energy for the backbone and side-chain atoms, as well as the solvent accessible areas of the corresponding beads. We then divided the all-atom desolvation energy value by the area of the bead, obtaining per-bead solvation parameters. For a comparison, all-atom and coarse-grained desolvation energy values of each residue of the protein barnase (PDB ID: 1A19, chain A) are shown mapped onto the structure in Figure 2A and 2B, respectively. The coarse-grained parameters for each residue (backbone and side-chain) are listed in Table 1.

Table 1. Coarse-grained solvation parameters (kCal.mol^{-1}.Å^{-2}).

Residue Name	Backbone	Side-Chain
Alanine	-0,0107	-
Glycine	-0,0089	-
Isoleucine	-0,0153	0,0255
Valine*	-0,0158	0,0222
Proline*	-0,0046	0,0230
Asparagine	-0,0137	-0,0192
Glutamine	-0,0147	-0,0135
Threonine	-0,0165	-0,0009
Serine	-0,0154	-0,0056
Methionine	-0,0130	0,0202
Cysteine	-0,0167	0,0201
Phenylalanine*	-0,0126	0,1005
Tyrosine*	-0,0134	0,0669
Tryptophan*	-0,0134	0,0872
Aspartatic Acid	-0,0169	-0,0360
Glutamic Acid	-0,0150	-0,0301
Histidine*	-0,0155	0,0501
Lysine*	-0,0163	-0,0210
Arginine*	-0,0162	-0,0229

*These residues have more than one bead representing their side-chain.

Validation of structure, restraint, and other input data

The success of HADDOCK is largely due to the availability of a web server interface that simplifies the modeling process and requires little to no technical knowledge[118]. Although the web server can expose several hundred parameters, depending on the access level of the user, some features require access to a local installation, such as custom ligand parameterization. The number of issues users often encounter using the local version of the software led us to implement several validation routines, some of them already present in the web server, directly into the software. These include, but are not limited to, validation of the correct file types for each input (structures, restraints, etc.), automatic generation of restraints to keep incomplete structures (i.e. with missing residues) or structures with multiple chains intact during the refinement process, and validation of the syntax in the restraint files. We also implemented a force field consistency check, to warn users of non-supported residues/molecules. We provide a wrapper script for the popular parameterization program PRODRG[234] and are planning on supporting the open-source ACPYPE program[235]. A similar wrapper is provided for the REDUCE, part of the Molprobity software[143], which can be used to automatically determine the protonation state of histidine side-chains and correct geometrical features of the input models. Finally, the coarse-grained conversion machinery is now integrated in the main software and requires no intervention from the users, except flagging which molecules should be converted.

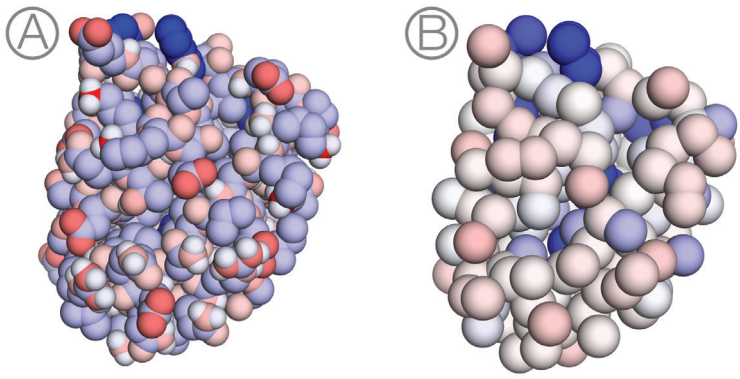


Figure 2. Solvation energy mapped to the structure of barnase (PDB Id: 1A19): **(A)** all atom structure and parameters, **(B)** the corresponding coarse grained structure and converted parameters. The colors show the range of solvation energy values: from negative (red) to positive (blue), for each particular atom/bead.

Testing the implementation of HADDOCK 3.0

To evaluate the performance of the above-mentioned modifications to HADDOCK, in particular the new orientation algorithm and the coarse-graining implementation, we collected a small dataset consisting of three multi-components protein complexes (Table 2). These tests are by no means extensive, only preliminary indicators of performance and quality. Further and more extensive testing must follow. We ran docking for all complexes using both HADDOCK 2.1 (requiring manual intervention to build the coarse-grained systems) and HADDOCK 3.0, except for heptameric 1K8K case, which could not run on HADDOCK 2.1 since the number of chains exceeded six. The same version of the CNS software (1.3) was used in all cases. For the rigid-body energy minimization, the sampling of 180º rotated solutions was disabled.

As a first test assuming ideal data, we extracted the minimum atom-atom distance for each residue pair in the interface of the crystal structures, up to a maximum of 5Å, and converted them to unambiguous distance restraints with a lower boundary equal to the target distance and an upper boundary of 0.05Å. Coarse-grained runs used the same data, except for the definition of the distance restraints between residues (instead of particular atoms) and having an upper boundary of 2.0Å to accommodate the difference in size and position of the beads. Wherever applicable, we imposed symmetry restraints to guide the conformational search, together with non-crystallographic symmetry restraints to force the conformations of the symmetrical molecules to remain as close as possible from each other.

Each run generated 1000/200/200 models, as default, at each respective stage. We kept default value since our restraint data was ideal; with experimental restraints, the number of models should be increased to assure adequate sampling. All other parameters not mentioned above were kept at default values.

Table 2. Dataset of protein complexes used to benchmark HADDOCK 3.0.

Structure Title	PDB Id	No. of Chains	Molecular Weight (kDa)	Symmetry
Vitamin B12 Transporter BtuCD-F	4FI3	5	157.7	A2B2C
Nicotinic Acetylcholine Receptor	2BG9	5	211.7	None
Arp2/3 Complex	1K8K	7	224.2	None

Results And Discussion

The sphere-based placement algorithm greatly improves the conformational search

In order to judge the performance of the sphere-based placement algorithm, we compared the number of models with an acceptable quality (near-native) generated by rigid-body energy minimization between HADDOCK 2.1 and HADDOCK 3.0. We define as acceptable quality a model with an RMSD from the crystal structure below 4Å, after fitting on all carbon-alpha atoms. The results are summarized in Table 3.

Despite perfect interface data, HADDOCK 2.1 generated very few acceptable solutions for the two pentameric complexes of the test dataset. For the vitamin transporter BtuCD-F (Figure 3A), despite the reduction of the conformational search space via symmetry restraints, the best model generated at the first stage is at 3.0Å RMSD, while the second best model is at 7.8Å. Yet, the 3Å model is the best scoring of the entire pool by a large margin (8 kcal.mol^{-1} vs 1062 kcal.mol^{-1} for the second best model), with the difference in energy stemming solely from the restraints energy term. This seems to indicate that the conformational landscape has a distinct minimum at the native conformation, but other local minima hinder the energy minimization protocol, which thus produces acceptable conformations only rarely (1 out of 5000 trials). The intertwining of the two ATP-binding proteins BtuC and BtuD (Figure 3A, bottom, green and yellow) is responsible for the high RMSD values, since visual inspection of the 'wrong' models shows the remaining three chains fitting the crystal almost perfectly. Nevertheless, the sphere-based placement algorithm bypasses this problem altogether, generating 716 models with acceptable quality, 666 of which below 1Å RMSD from the crystal structure. The initial spacing calculated by HADDOCK 3.0 is set at 95Å, compared to the default 50Å in HADDOCK 2.1, which might explain the difference in quality of the models. With a larger separation, the molecules have more chances to feel and adapt to each other. However, increasing the separation distance in HADDOCK 2.1 to 100Å does not improve the results (data not shown). Another explanation is the coordinates of the initial positions of the molecules. HADDOCK 2.1 places the molecules at five vertices of an octahedron, which is the optimal shape for six molecules, and therefore creates a rather asymmetric distribution. The sphere-based algorithm provides the minimal energy configuration for exactly five molecules that is apparently more adequate.

The nicotinic acetylcholine receptor complex (Figure 3B) shows again the sphere-based implementation performing best. HADDOCK 3.0 is superior, generating 8 models vs. 0 models by HADDOCK 2.1 with an RMSD below 4Å. The generation of so few acceptable models can be partially attributed to the 1) the small number of models requested at *it0* (1000), 2) the lack of symmetry restraints, as opposed to the previous case, that are of substantial help in reducing the conformational/interaction space. In addition, the size of this system (211,7 kDa, 1849 residues, 5 chains) might complicate the energy landscape and

therefore the execution of the minimization algorithm. Since HADDOCK was designed primarily for dimers, some parameters, such as the number of steps in the minimization algorithm, are probably not optimal as the system size grows. We found that for both versions of HADDOCK, the minimization algorithm reached the maximum number of steps before converging, indicating that the minimization parameters/protocol could still be improved. Doubling the number of steps in the various stages of the rigid-body energy minimization (from 250 to 500) greatly increased the number of models of acceptable quality (HADDOCK 3.0 to 438; HADDOCK 2.1 to 11) but the minimization itself still failed to reach convergence. As a test, we continuously increased the number of steps until the minimization algorithm reached the convergence criterion (norm of the gradient of total energy below 0.0001, as default in CNS), which happened at ~3000 steps (12 times the original value). However, at least for this particular case, the number of acceptable quality models did not increase linearly with a much larger number of steps performed by the minimization algorithm (687 models of acceptable quality). Regardless of the number of steps, the best-ranked models according to the HADDOCK score are those showing the best resemblance to the native structure.

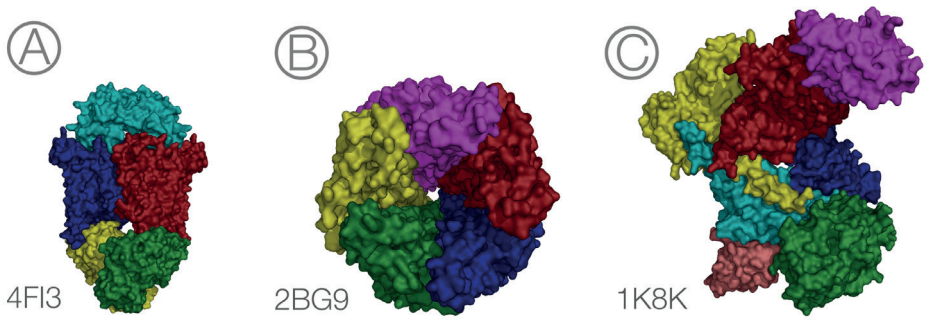


Figure 3. The three protein-protein complexes used as a test set for HADDOCK 3.0. The complexes are ordered, from left to right, by increasing molecular weight and number of chains. Each chain is coloured individually.

The third and final protein-protein complex of the dataset is the Arp2/3 complex (Figure 3C), an asymmetric heptamer with very complicated geometry: it has both large and small interface areas (respectively, chains D/E, yellow and cyan, and chains A/F, red and purple) and an intertwining of chains A/D/E that comprise the core of the complex. HADDOCK 2.1, being limited to six molecules, could not run for this case. Based on the experience of the previous case, we ran calculations with both the default number of steps for the energy minimization and the double figures. The results are quite satisfactory: with the original number of EM steps (250), 7 models are within 4Å RMSD of the crystal structure, while the extended minimization protocol (500 steps) produces 137 acceptable models within the 1000 written to disk. In addition, for this complex, we again increased the number of steps until we observed convergence of the minimization algorithm, which happened at ~2500 steps. Much like with 2BG9, this did not produce a significantly larger number of acceptable

quality models (241). As with the previous targets, the best models in terms of HADDOCK score correspond to native-like conformation.

These results indicate that for larger molecules, the rigid-body protocol benefits from an adjustment to the number of steps performed by the Powell minimizer. Increasing the number of steps is not significantly degrading the performance of HADDOCK, and for smaller systems, the minimizer is anyway likely to converge before reaching the maximum number of steps. It might be interesting to implement a number of steps that depends on the number of molecules being docked. However, it is worth noting that, despite doubling the number of steps, the minimizer still fails to converge in any of the cases. Yet, since the algorithm already generates a large number of acceptable models, and following the observations for 2BG9 and 1K8K, it might be unnecessary to increase the number of steps to reach convergence, under penalty of increasing the calculation time substantially. It is thus probably best to revisit the convergence criterion, which might be too strict.

Coarse-graining reduces the calculation time dramatically while keeping reasonable modeling accuracy

The reasoning behind coarse-grained modeling was to reduce the calculation time needed for each stage, namely the computationally expensive semi-flexible refinement. However, coarse-graining reduces the details of the structure representation and we could expect a drop in the quality of the models. This quality was evaluated similarly to that of the all-atom models, but using as reference the coarse-grained conversion of the crystal structure, and fitting and calculating the RMSD on the BB beads.

To our surprise, for the vitamin transporter BtuCD-F, the coarse-grained protocol produced a significant number of acceptable models (694), even when using the original placement algorithm of HADDOCK 2.1 (40). This highlights an attractive feature of coarse-grained modeling: the less detailed models produce a smoother energy landscape that the minimizer can navigate more efficiently. The performance improvement was also significant: 336 seconds vs. 1150 for the combined rigid-body and semi-flexible refinement steps. The semi-flexible refinement step, in particular, takes on average one-fifth of the time when compared to the regular all-atom protocol. The conversion to atomic coordinates seems to be, in fact, the slowest step, which might be due to the several rounds of optimization and refinement. In addition, as expected, water refinement after the all-atom conversion is computationally demanding. The results for the two other complexes are similar: a dramatic reduction in computation time coupled to a slight loss of accuracy in the produced models and an increase of the number of acceptable solutions with the extended minimization algorithm. For the nicotinic acetylcholine receptor, coarse-graining takes 896 seconds vs. 2445 seconds for the combined *it0* + *it1* refinement with the coarse-grained model and the all-atom representation, respectively. The number of acceptable quality models at the rigid body stage drops by ~60% (CG: 176 vs. AA: 438, out of 1000 written to disk) although these are still top ranking according to the HADDOCK score. The Arp2/Arp3 heptamer complex failed to produce acceptable models when running with the coarse-grained representation, which was expected since the all-atom calculations only produced 7 acceptable models. As before, doubling the number of steps in the minimization algorithm during the rigid-body stage resulted in the generation of more acceptable models (15 coarse-grained vs. 137 all-atom) at negligible computational cost. The time saved by coarse-graining could then be invested in more extensive sampling by increasing the number of models generated, or increasing the length of the refinement stages (i.e. number of steps).

Table 3. Modelling of the three protein-protein complexes of the dataset for both versions of HADDOCK, representations (AA and CG) and number of steps during it0 minimization (250 and 500). The time spent for the water refinement stage of coarse-grained runs includes the morphing process.

Structure	HADDOCK version	Resolution	Steps in it0	Average Time (s)*			No. of acceptable models in it0	RMSD of best model in it0 (Å)
				it0	it1	itw		
Vitamin B12 Transporter BtuCD-F	2.1	AA	250	138	1012	1233	1	3.0
	3.0			148	1030	1108	716	0.2
	2.1	CG		91	263	1701	40	1.6
	3.0			81	255	1689	694	1.0
Nicotinic Acetylcholine Receptor	2.1	AA	250	621	1493	3293	0	17.1
			500	876	1547	2898	11	0.2
	3.0		250	614	1831	3727	8	0.2
			500	888	1587	2540	438	0.2
	2.1	CG	250	454	483	3771	0	4.3
			500	610	512	2870	213	3.2
	3.0		250	471	425	3078	176	3.2
			500	638	434	3376	874	3.2
Arp2/3 Complex	3.0	AA	250	606	1574	2511	7	1.8
			500	889	1780	2567	137	0.2
		CG	250	472	472	3323	0	4.1
			500	659	472	3842	15	3.0

*This value is an average of the time reported by CNS in each individual calculation log file. Each calculation is the combination of ten trials (5 rigid-body energy minimization round + 180-degree rotation sampling for each). The calculations were performed on a Linux cluster with AMD Opteron 6344 processors.

To assess the individual quality of the models, we compared the best model generated by rigid-body minimization, judged by RMSD from the native structure, of both all-atom and coarse-grained docking. Overall, the all-atom models are more accurate, which is natural due to the loss of detail when using a coarse-grained representation. Nevertheless, in some cases, particularly when the all-atom calculations failed due to the low number of steps of the minimization algorithm, coarse-graining produced surprisingly good models. The reason might be, as previously mentioned, that the less detailed models result in a smoother energy landscape, which the minimizer is able to traverse more easily. This is evident in the coarse-grained modeling of 4FI3 and 2BG9 using HADDOCK 2.1, which despite departing from non-optimal initial orientations, produced models of much better quality (1.6Å vs. 3.0Å for 4FI3; 4.3Å vs. 17.1Å for 2BG9). This highlights the benefits of using a coarse-grained representation for very large systems, in particular for the initial conformational search stage.

Conclusions

Algorithms that support an arbitrary number of molecules will be a vital determinant for the survival of software packages in order to cope with the increasing complexity of the systems being studied in (computational) structural biology. The gradual shift to larger and more complex molecular assemblies also imposes a strain on the computational efficiency. In

this work, we proposed two modifications to the HADDOCK integrative modeling software: an initial placement algorithm generalizable to an arbitrary number of molecules and the integration of a coarse-grained force field in its workflow. The preliminary results on a small dataset of protein-protein complexes indicate that these features improve both the performance and the efficiency of the docking calculations. Further testing is obviously required in order to cement the conclusions of this work, preferably on a larger dataset of diverse biological systems and without the bias of perfect interface data. It is also important to investigate the impact of these modifications on more simple systems, such as binary complexes, and define the limits of their applicability. Another interesting research line is the usability of different types of restraints in modeling large complexes. As seen from the results, even with perfect unambiguous distance restraints there are still quite some false positive models. Fortunately, the scoring function of HADDOCK seems to be robust enough to detect these cases and select near-native conformations, for both all-atom and coarse-grained representations.

Acknowledgements

Ms. Charleen Don, as part of her M.Sc dissertation, carried out the MARTINI coarse-grained force field implementation and optimization of the desolvation energy parameters. We also acknowledge the developers of the MARTINI force field for the availability of their tools, in particular the conversion script *martinize.py*, on which we based on conversion machinery.

Conclusions & Perspectives

"It's is not bad, but it isn't prize-winning either..."

- dr. Markus Weingarth

On the verge of half a century of continuous developments, Computational Structural Biology is a mature field of research with both academic and industrial applications. Driven by the needs and discoveries of experimental structural biology and biophysics, *in silico* research remains daring and innovative, expanding its reach with every new technological and/or scientific breakthrough.

When I joined the Computational Structural Biology group at Utrecht University, in November of 2008 as a master student, the longest molecular dynamics effort to date had achieved 10 μs of simulated time[236]. In the field of protein-protein docking, HADDOCK had established itself as a leading method according to the CAPRI assessment panel[11]. When I started my Ph.D. in 2010, David Shaw had just published in Science the first *millisecond* molecular dynamics simulation of protein folding[237] and HADDOCK had been crowned the top-ranking docking method at CAPRI[112] and was the most cited docking program for structure prediction of interactions[57]. In addition, a number of developments were in progress in the group, such as multibody docking[69], new experimental restraints[66], and advances in understanding the roots of binding affinity and specificity in protein-protein interactions[103,238].

In the following years, I was fortunate to witness the development of robust approaches for protein-protein interaction design[239], some of which produced targets we worked on - and successfully predicted - for CAPRI[125], and the expansion of HADDOCK to new, up-to-then unexplored, territories (very-large conformational changes[16], protein-peptide docking[70]). During this time, the field was democratized through the development of interactive computer games, such as FoldIt[240] and UDOCK[241], which not only allow every individual to join in structure prediction, but also provide outside-the-box solutions to traditional problems[242,243]. Finally, one year ago, just as this thesis was starting to be organized, the Royal Swedish Academy of Sciences awarded the Nobel Prize in Chemistry to key figures of our field: Martin Karplus, Michael Levitt, and Arieh Warshel.

The contributions laid out in this thesis have addressed important open problems in the field of structure prediction of protein interactions: large biological assemblies, interface prediction, and the improvement of sampling and scoring through integrative modeling. Every chapter here presented has built on, drawn inspiration from, or followed the many previously mentioned developments and breakthroughs, in particular those of the HADDOCK research group.

Chapter 2 analyzed a common practice during CAPRI predictions, and docking in general, which is the creation of homology models for proteins that have no experimentally determined structure but that participate in a complex whose structure we aim to predict. We have shown that among several structure- and sequence-based metrics, the sequence identity between the target protein and the template is the best and simplest indicator of success for the docking predictions. However, a *de facto* rule-of-thumb in homology modeling, the dependence of the quality of the final model on the quality of the target/template sequence alignment, was not covered by our study and should be addressed in the future.

Chapter 3 presented one of the first applications of protein coevolution analysis to the problem of protein-protein interface prediction. Despite producing extremely accurate predictions for several cases, the applicability of such coevolution methods remains limited for the wider structural biology community. First, to derive robust residue-residue correlations the current implementation of the methods require an abundance of sequence information, in the form of large (and diverse) multiple sequence alignments, which are not always available, in particular in evolutionarily young protein families. Second, specifically

for residue correlations *between* interacting proteins, the multiple sequence alignments must contain paired sequences of interacting pairs of proteins, which is problematic due to speciation and duplication events. While in prokaryotes this pairing problem is minimized by relying on the operon-based genome organization, there is no corresponding strategy for eukaryotes. Applications to many important systems for human biology remain therefore out of reach.

The last two chapters address challenges posed by large and complex molecular assemblies both to the computational efficiency and to the accuracy of docking methods, which became more apparent in recent years due to the increased experimental focus on these systems. **Chapter 4** introduced a contact-based similarity measure for protein-protein docking models that is two orders of magnitude faster than traditional methods, and even faster when applied to symmetrical assemblies. Its main advantages are that it does not require fitting of the models and that it can handle an arbitrary number of components at minimal cost. The resulting clustering methodology is to become default in future versions of HADDOCK. **Chapter 5** revisited the underlying workflow and parameters of HADDOCK, implementing a new pre-docking orientation protocol and the MARTINI coarse-grained force field. The results show improvements in the quality of the models, as well as a significant reduction in computation time, while maintaining reasonable quality, when using the reduced representation of MARTINI. These results remain preliminary until testing on a larger and more diverse collection of structures.

Overall, the computational structural biology of macromolecular interactions is a thriving field of research that bridges with several others such as molecular simulation, structural biophysics, and molecular biology. However, and as it must happen, scientific progress not only provides answers but it also raises **new questions**.

CPORT is an interface prediction method that combines predictions from up to six different methods, producing a consensus prediction[88]. It provides ambiguous interface information and sacrifices specificity (fraction of true positive predictions) for high sensitivity (fraction of the interface that is correctly predicted), which results, more often than not, in several large predicted interfaces. HADDOCK deals with such predicament by removing, at random, a large percentage (87.5%) of the prediction-based restraints for each docking trial. Consequently, docking based on CPORT predictions tends to produce many clusters, meaning that the onus is then on the (imperfect) scoring function. A possible strategy to produce very high-quality docking models from interface prediction data only is the combination of methods such as CPORT with those presented in **Chapter 4**. On the one hand, the coevolution predictions can help disambiguate results from traditional predictors; conversely, the information used by these predictors (e.g. amino acid interface propensity and physical characteristics) can be used to re-weight the coevolving residue pairs and thus increase the discrimination of false-positives.

As observed in **Chapter 6**, the number of high-resolution distance restraints can contribute negatively to the convergence of the minimization algorithms of HADDOCK. This is particularly important for large biological systems, as they require an equally large number of restraints to accurately describe the relative orientation and distance of each molecule to (all) others. Alternatively, low-resolution shape information from cryo-electron microscopy can be processed very efficiently to provide initial positions for docking calculations[14,244], in particular for large biological assemblies. The combination of such initial restraints with flexible refinement driven by distance restraints, as implemented in HADDOCK, should yield an extremely powerful method that leverages both ends of the resolution scale to

model complexes at atomic detail.

Finally, the elephant in the room is clearly the problem of docking proteins *within* a lipid bilayer and/or proteins *to* a lipid bilayer. Although nearly a third of the proteins encoded in most genomes interact somehow with a biological membrane[245], except for a few isolated applications[246], none of the top-ranked methods in CAPRI has a straightforward membrane-dedicated approach. This is understandable since the membrane environment is a remarkably thorny biophysical system. On the other hand, there have been enough advances in computational and experimental methods[247-249] to ease the development of (data-driven) docking methods for membrane proteins. This raises the question: with so many technological advances and experimental data available, what is the docking community waiting for to address the fraction of the protein universe targeted by *more than half* of the marketed medicinal drugs[250]? An often-heard justification is the lack of experimentally determined structures of integral membrane proteins, although recent statistics[251] and developments in structure prediction methods[167,252,253] hint at the contrary. Also contributing to the development of membrane-aware docking protocols will be the membrane itself, which imposes spatial and chemical restraints on the conformation and orientation of proteins. Further, interface prediction methods such as those outlined in **Chapter 4** do not rely on structural data and as such are readily applicable to membrane-related protein complexes of unknown structure.

As a final conclusion, the challenges tackled by the body of work in this thesis, as well as those outlined in the previous paragraphs, are only a small fraction of what the field of computational structural biology faces in order to remain state-of-the-art. Open problems will remain, such as dealing with intrinsically disordered domains/proteins, describing folding upon binding, and predicting interaction specificity. Fortunately, the last six years have demonstrated that this field of research is surprisingly dynamic and innovative, and most importantly, not afraid to test its limits.

References

1. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome Networks and Human Disease. *Cell* **144,** 986–998 (2011).
2. Eichborn, Von, J., Günther, S. & Preissner, R. Structural features and evolution of protein-protein interactions. *Genome Inform* **22,** 1–10 (2010).
3. Chruszcz, M., Domagalski, M., Osinski, T., Wlodawer, A. & Minor, W. Unmet challenges of structural genomics. *Curr Opin Struct Biol* 1–11 (2010). doi:10.1016/j.sbi.2010.08.001
4. Levitt, M. Nature of the protein universe. *Proceedings of the National Academy of Sciences* **106,** 11079 (2009).
5. Kolodny, R., Pereyaslavets, L., Samson, A. O. & Levitt, M. On the Universe of Protein Folds. *Annu Rev Biophys* **42,** 130325113559002 (2012).
6. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10,** 47–53 (2013).
7. Levitt, M. The birth of computational structural biology. *Nature Structural & Molecular Biology* **8,** 392–393 (2001).
8. Schlick, T., Collepardo-Guevara, R., Halvorsen, L. A., Jung, S. & Xiao, X. Biomolecular modeling and simulation: a field coming of age. *Quarterly Reviews of Biophysics* **44,** 191–228 (2011).
9. Janin, J. *et al.* CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Bioinformatics* **52,** 2–9 (2003).
10. Méndez, R., Leplae, R., Lensink, M. F. & Wodak, S. J. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins: Structure, Function, and Bioinformatics* **60,** 150–169 (2005).
11. Lensink, M. F., Méndez, R. & Wodak, S. J. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins: Structure, Function, and Bioinformatics* **69,** 704–718 (2007).
12. Lensink, M. F. & Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. *Proteins: Structure, Function, and Bioinformatics* **78,** 3073–3084 (2010).
13. Lensink, M. F. & Wodak, S. J. Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics* **81,** 2082–2095 (2013).
14. Karaca, E. & Bonvin, A. M. J. J. Advances in integrative modeling of biomolecular complexes. *Methods* **59,** 372–381 (2013).
15. Rodrigues, J. P. G. L. M. *et al.* Defining the limits of homology modelling in information-driven protein docking. *Proteins: Structure, Function, and Bioinformatics* **81,** 2119–2128 (2013).
16. Karaca, E. & Bonvin, A. M. J. J. A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. *Structure* **19,** 555–565 (2011).
17. Kastritis, P. L. & Bonvin, A. M. Molecular origins of binding affinity: seeking the Archimedean point. *Curr Opin Struct Biol* (2013). doi:10.1016/j.sbi.2013.07.001
18. Kastritis, P. L. & Bonvin, A. M. J. J. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J R Soc Interface* **10,** 20120835 (2013).
19. Vajda, S. & Kozakov, D. Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol* **19,** 164–170 (2009).
20. Katchalski-Katzir, E. *et al.* Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci USA* **89,** 2195–2199 (1992).
21. Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins: Structure, Function, and Bioinformatics* **52,** 80–87 (2003).
22. Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **20,** 45–50 (2004).
23. Cheng, T. M.-K., Blundell, T. L. & Fernández-Recio, J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins: Structure, Function, and Bioinformatics* **68,** 503–515 (2007).
24. Fischer, D., Bachar, O., Nussinov, R. & Wolfson, H. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J. Biomol. Struct. Dyn.* **9,** 769–789 (1992).
25. Mashiach, E. *et al.* An integrated suite of fast docking algorithms. *Proteins: Structure, Function, and Bioinformatics* **78,** 3197–3204 (2010).
26. Ritchie, D. W. & Kemp, G. J. Protein docking using spherical polar Fourier correlations. *Proteins: Structure, Function, and Bioinformatics* **39,** 178–194 (2000).

27. Macindoe, G., MAVRIDIS, L., VENKATRAMAN, V., Devignes, M. D. & Ritchie, D. W. HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res* **38,** W445–W449 (2010).

28. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics* **65,** 392–406 (2006).

29. Vreven, T., Hwang, H. & Weng, Z. Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci* **20,** 1576–1586 (2011).

30. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125,** 1731–1737 (2003).

31. Moal, I. H. & Bates, P. A. SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *Int J Mol Sci* **11,** 3623–3648 (2010).

32. Zacharias, M. ATTRACT: protein-protein docking in CAPRI using a reduced protein model. *Proteins: Structure, Function, and Bioinformatics* **60,** 252–256 (2005).

33. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci USA* **44,** 98–104 (1958).

34. Bonvin, A. M. Flexible protein–protein docking. *Curr Opin Struct Biol* **16,** 194–200 (2006).

35. Zacharias, M. Accounting for conformational changes during protein-protein docking. *Curr Opin Struct Biol* **20,** 180–186 (2010).

36. Gray, J. J. *et al.* Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* **331,** 281–299 (2003).

37. May, A. & Zacharias, M. Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochim Biophys Acta* **1754,** 225–231 (2005).

38. de Vries, S. & Zacharias, M. Flexible docking and refinement with a coarse-grained protein model using ATTRACT. *Proteins: Structure, Function, and Bioinformatics* (2013). doi:10.1002/prot.24400

39. Moal, I. H., Moretti, R., Baker, D. & Fernández-Recio, J. Scoring functions for protein-protein interactions. *Curr Opin Struct Biol* (2013). doi:10.1016/j.sbi.2013.06.017

40. Champ, P. C. & Camacho, C. J. FastContact: a free energy scoring tool for protein-protein complex structures. *Nucleic Acids Res* **35,** W556–60 (2007).

41. Andrusier, N., Nussinov, R. & Wolfson, H. J. FireDock: fast interaction refinement in molecular docking. *Proteins: Structure, Function, and Bioinformatics* **69,** 139–159 (2007).

42. Moal, I. H., Torchala, M., Bates, P. A. & Fernández-Recio, J. The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinformatics* **14,** 286 (2013).

43. Pons, C., Talavera, D., la Cruz, de, X., Orozco, M. & Fernández-Recio, J. Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. *J Chem Inf Model* **51,** 370–377 (2011).

44. Liu, S. & Vakser, I. A. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC Bioinformatics* **12,** 280 (2011).

45. Viswanath, S., Ravikant, D. V. S. & Elber, R. Improving ranking of models for protein complexes with side chain modeling and atomic potentials. *Proteins: Structure, Function, and Bioinformatics* **81,** 592–606 (2013).

46. Pierce, B. & Weng, Z. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins: Structure, Function, and Bioinformatics* **67,** 1078–1086 (2007).

47. Liu, S., Zhang, C., Zhou, H. & Zhou, Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics* **56,** 93–101 (2004).

48. Khashan, R., Zheng, W. & Tropsha, A. Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins: Structure, Function, and Bioinformatics* **80,** 2207–2217 (2012).

49. Andreani, J., Faure, G. & Guerois, R. InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics* **29,** 1742–1749 (2013).

50. Tress, M. *et al.* Scoring docking models with evolutionary information. *Proteins: Structure, Function, and Bioinformatics* **60,** 275–280 (2005).

51. Vangone, A., Cavallo, L. & Oliva, R. Using a consensus approach based on the conservation of inter-residue contacts to rank CAPRI models. *Proteins: Structure, Function, and Bioinformatics* (2013). doi:10.1002/prot.24423

52. Shortle, D., Simons, K. T. & Baker, D. Clustering of low-energy conformations near the native structures of small proteins. *Proc Natl Acad Sci USA* **95,** 11158–11162 (1998).

53. Camacho, C. J., Weng, Z., Vajda, S. & DeLisi, C. Free energy landscapes of encounter complexes in protein-protein association. *Biophys J* **76,** 1166–1178 (1999).

54. Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res* **32,** W96–9 (2004).

55. Vreven, T., Pierce, B. G., Hwang, H. & Weng, Z. Performance of ZDOCK in CAPRI rounds 20-26. *Proteins: Structure, Function, and Bioinformatics* **81,** 2175–2182 (2013).

56. Rodrigues, J. P. G. L. M. *et al.* Clustering biomolecular complexes by residue contacts similarity. *Proteins: Structure, Function, and Bioinformatics* **80,** 1810–1817 (2012).

57. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Protein-protein docking dealing with the unknown. *J Comput Chem* NA–NA (2009). doi:10.1002/jcc.21276

58. van Dijk, A. D. J., Boelens, R. & Bonvin, A. M. J. J. Data-driven docking for the study of biomolecular complexes. *FEBS Journal* **272,** 293–312 (2004).

59. de Vries, S. J. *et al.* Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. *Proteins: Structure, Function, and Bioinformatics* **78,** 3242–3249 (2010).

60. Tovchigrechko, A. & Vakser, I. A. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res* **34,** W310–4 (2006).

61. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOL– a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J Appl Crystallogr* **28,** 768–773 (1995).

62. Pons, C. *et al.* Structural characterization of protein-protein complexes by integrating computational docking with small-angle scattering data. *J Mol Biol* **403,** 217–230 (2010).

63. Chelliah, V., Blundell, T. L. & Fernández-Recio, J. Efficient Restraints for Protein–Protein Docking by Comparison of Observed Amino Acid Substitution Patterns with those Predicted from Local Environment. *J Mol Biol* **357,** 1669–1682 (2006).

64. Brunger, A. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54,** 905–921 (1998).

65. van Dijk, A. D. J., Fushman, D. & Bonvin, A. M. J. J. 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: Structure, Function, and Bioinformatics* **60,** 367–381 (2005).

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

67. Karaca, E. & Bonvin, A. M. J. J. On the usefulness of ion-mobility mass spectrometry and SAXS data in scoring docking decoys. *Acta Cryst (2013). D69, 683-694 [doi:10.1107/S0907444913007063]* 1–12 (2013). doi:10.1107/S0907444913007063

68. Nilges, M., Gronenborn, A. M., Brünger, A. T. & Clore, G. M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng* **2,** 27–38 (1988).

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

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

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

72. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* **33,** W363–7 (2005).

73. Mashiach-Farkash, E., Nussinov, R. & Wolfson, H. J. SymmRef: a flexible refinement method for symmetric multimers. *Proteins: Structure, Function, and Bioinformatics* **79,** 2607–2623 (2011).

74. Tjioe, E., Lasker, K., Webb, B., Wolfson, H. J. & Sali, A. MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map. *Nucleic Acids Res* **39,** W167–70 (2011).

75. Ritchie, D., VENKATRAMAN, V. & MAVRIDIS, L. Using Graphics Processors to Accelerate Protein Docking

Calculations. *Studies in health technology and informatics* **159,** 146 (2010).

76. Pierce, B. G., Hourai, Y. & Weng, Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS ONE* **6,** e24657 (2011).

77. Choi, U. B. *et al.* Single-molecule FRET-derived model of the synaptotagmin 1-SNARE fusion complex. *Nature Publishing Group* **17,** 318–324 (2010).

78. Kalisman, N., Adams, C. M. & Levitt, M. Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. *Proc Natl Acad Sci USA* **109,** 2884–2889 (2012).

79. Feng, C. Mechanism of Nitric Oxide Synthase Regulation: Electron Transfer and Interdomain Interactions. *Coordination Chemistry Reviews* **256,** 393–411 (2012).

80. Hennig, J., Wang, I., Sonntag, M., Gabel, F. & Sattler, M. Combining NMR and small angle X-ray and neutron scattering in the structural analysis of a ternary protein-RNA complex. *J Biomol NMR* **56,** 17–30 (2013).

81. Uetrecht, C., Rose, R. J., van Duijn, E., Lorenzen, K. & Heck, A. J. R. Ion mobility mass spectrometry of proteins and protein assemblies. *Chem Soc Rev* **39,** 1633–1655 (2010).

82. de Vries, S. J., van Dijk, A. D. J. & Bonvin, A. M. J. J. WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins: Structure, Function, and Bioinformatics* **63,** 479–489 (2006).

83. Zhang, Q. C. *et al.* PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res* **39,** W283–W287 (2011).

84. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38,** W529–33 (2010).

85. Negi, S. S., Schein, C. H., Oezguen, N., Power, T. D. & Braun, W. InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics* **23,** 3397–3399 (2007).

86. Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins: Structure, Function, and Bioinformatics* **66,** 630–645 (2007).

87. Neuvirth, H., Raz, R. & Schreiber, G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol* **338,** 181–199 (2004).

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

89. Zellner, H. *et al.* PresCont: predicting protein-protein interfaces utilizing four residue properties. *Proteins: Structure, Function, and Bioinformatics* **80,** 154–168 (2012).

90. Qin, S. & Zhou, H.-X. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics* **23,** 3386–3387 (2007).

91. Ahmad, S. & Mizuguchi, K. Partner-Aware Prediction of Interacting Residues in Protein-Protein Complexes from Sequence Data. *PLoS ONE* **6,** e29104 (2011).

92. Xue, L. C., Dobbs, D. & Honavar, V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics* **12,** 244 (2011).

93. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* **106,** 67–72 (2009).

94. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat Rev Genet* **14,** 249–261 (2013).

95. Hwang, H., Vreven, T. & Weng, Z. Binding interface prediction by combining protein-protein docking results. *Proteins: Structure, Function, and Bioinformatics* (2013). doi:10.1002/prot.24354

96. Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10,** e1001244 (2012).

97. Aytuna, A. S., Gursoy, A. & Keskin, O. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* **21,** 2850–2855 (2005).

98. Mukherjee, S. & Zhang, Y. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure* **19,** 955–966 (2011).

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

100. Vreven, T., Hwang, H., Pierce, B. G. & Weng, Z. Evaluating template-based and template-free protein-protein complex structure prediction. *Brief. Bioinformatics* **15,** 169–176 (2013).

101. Ghoorah, A. W., Devignes, M.-D., Smaïl-Tabbone, M. & Ritchie, D. W. Protein docking using case-based reasoning. *Proteins: Structure, Function, and Bioinformatics* (2013). doi:10.1002/prot.24433

102. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. The Relationship Between Sequence and Interaction Divergence in Proteins. *J Mol Biol* **332,** 989–998 (2003).

103. van Wijk, S. J. L., Melquiond, A. S. J., de Vries, S. J., Timmers, H. T. M. & Bonvin, A. M. J. J. Dynamic Control of Selectivity in the Ubiquitination Pathway Revealed by an ASP to GLU Substitution in an Intra-Molecular Salt-Bridge Network. *PLoS Comput Biol* **8,** e1002754 (2012).

104. Liger, D. *et al.* Mechanism of activation of methyltransferases involved in translation by the Trm112 'hub' protein. *Nucleic Acids Res* **39,** 6249–6259 (2011).

105. Wodak, S. J. & Janin, J. Computer analysis of protein-protein interaction. *J Mol Biol* **124,** 323–342 (1978).

106. Melquiond, A. S., Karaca, E., Kastritis, P. L. & Bonvin, A. M. Next challenges in protein–protein docking: from proteome to interactome and beyond. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2,** 642–651 (2012).

107. de Vries, S. J. & Bonvin, A. M. J. J. How proteins get in touch: interface prediction in the study of biomolecular complexes. *Curr Protein Pept Sci* **9,** 394–406 (2008).

108. Ezkurdia, I. *et al.* Progress and challenges in predicting protein-protein interaction sites. *Brief. Bioinformatics* **10,** 233–246 (2008).

109. Alberts, B. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell* **92,** 291–294 (1998).

110. Levy, E. D. & Pereira-Leal, J. B. Evolution and dynamics of protein interactions and networks. *Curr Opin Struct Biol* **18,** 349–357 (2008).

111. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. Protein repeats: structures, functions, and evolution. *J Struct Biol* **134,** 117–131 (2001).

112. Lensink, M. F. & Wodak, S. J. Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins: Structure, Function, and Bioinformatics* **78,** 3085–3095 (2010).

113. Montelione, G. T. The Protein Structure Initiative: achievements and visions for the future. *F1000 Biol Rep* **4,** 7 (2012).

114. Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. Fast and accurate automatic structure prediction with HHpred. *Proteins: Structure, Function, and Bioinformatics* **77 Suppl 9,** 128–132 (2009).

115. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5,** 725–738 (2010).

116. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12,** 85–94 (1999).

117. de Vries, S. J. *et al.* HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins: Structure, Function, and Bioinformatics* **69,** 726–733 (2007).

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

119. Carvalho, A. L. *et al.* Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proc Natl Acad Sci USA* **100,** 13809–13814 (2003).

120. Sansen, S. *et al.* Structural basis for inhibition of Aspergillus niger xylanase by triticum aestivum xylanase inhibitor-I. *J Biol Chem* **279,** 36022–36028 (2004).

121. Bonsor, D. A., Grishkovskaya, I., Dodson, E. J. & Kleanthous, C. Molecular mimicry enables competitive recruitment by a natively disordered protein. *J Am Chem Soc* **129,** 4800–4807 (2007).

122. Walker, J. R. *et al.* Novel and Unexpected Complex Between the SUMO-1-Conjugating Enzyme UBC9 and the Ubiquitin-Conjugating Enzyme E2-25 kDa.

123. Bao, R. *et al.* The ternary structure of the double-headed arrowhead protease inhibitor API-A complexed with two trypsins reveals a novel reactive site conformation. *Journal of Biological Chemistry* **284,** 26676–26684 (2009).

124. Meenan, N. A. G. *et al.* The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc Natl Acad Sci USA* **107,** 10080–10085 (2010).

125. Fleishman, S. J. *et al.* Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332,** 816–821 (2011).

126. Minard, P. & Graille, M. Designed Rep4/Rep2 α-repeat complex.

127. van Dijk, A. D. J. *et al.* Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins: Structure, Function, and Bioinformatics* **60,** 232–238 (2005).

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

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

130. Bordogna, A., Pandini, A. & Bonati, L. Predicting the accuracy of protein-ligand docking on homology models. *J Comput Chem* **32,** 81–98 (2011).

131. Eramian, D., Eswar, N., Shen, M.-Y. & Sali, A. How well can the accuracy of comparative protein structure models be predicted? *Protein Sci* **17,** 1881–1893 (2008).

132. Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27,** 343–350 (2011).

133. Eisenberg, D., Lüthy, R. & Bowie, J. U. VERIFY3D: assessment of protein models with three-dimensional profiles. *Meth. Enzymol.* **277,** 396–404 (1997).

134. Wojdyla, J. A., Fleishman, S. J., Baker, D. & Kleanthous, C. Structure of the ultra-high-affinity colicin E2 DNase--Im2 complex. *J Mol Biol* **417,** 79–94 (2012).

135. Fox, B., Bailey, L. & Acheson, J. T4moC/T4moH mono-oxygenase complex.

136. Najmudin, S. *et al.* Putting an N-terminal end to the Clostridium thermocellum xylanase Xyn10B story: crystal structure of the CBM22-1-GH10 modules complexed with xylohexaose. *J Struct Biol* **172,** 353–362 (2010).

137. Minard, P. & Graille, M. Designed neocarzinostatin/Rep16 α-repeat complex.

138. Basle, A. & Lewis, R. A protein-polysaccharide complex.

139. Leysen, S., Vanderkelen, L., Weeks, S. D., Michiels, C. W. & Strelkov, S. V. Structural basis of bacterial defense against g-type lysozyme-based innate immunity. *Cell Mol Life Sci* **70,** 1113–1122 (2013).

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

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

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

143. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66,** 12–21 (2010).

144. Rodrigues, J. P. G. L. M., Levitt, M. & Chopra, G. KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Res* **40,** W323–8 (2012).

145. Boratyn, G. M. *et al.* BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* (2013). doi:10.1093/nar/gkt282

146. Holm, L., Kääriäinen, S., Rosenström, P. & Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24,** 2780–2781 (2008).

147. Taly, J.-F. *et al.* Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. *Nat Protoc* **6,** 1669–1682 (2011).

148. Simossis, V. A. & Heringa, J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res* **33,** W289–94 (2005).

149. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234,** 779–815 (1993).

150. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15,** 2507–2524 (2006).

151. Jorgensen, W. & Tirado-Rives, J. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* **110,** 1657–1666 (1988).

152. Linge, J. P., Habeck, M., Rieping, W. & Nilges, M. ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19,** 315–316 (2003).

153. Das, K. *et al.* Crystal structure of RlmAI: implications for understanding the 23S rRNA G745/G748-methylation at the macrolide antibiotic-binding site. *Proc Natl Acad Sci USA* **101,** 4041–4046 (2004).

154. Senda, M. *et al.* Molecular mechanism of the redox-dependent interaction between NADH-dependent ferredoxin reductase and Rieske-type [2Fe-2S] ferredoxin. *J Mol Biol* **373,** 382–400 (2007).

155. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31,** 3370–3374 (2003).

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

157. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat Biotechnol* **30,** 1072–1080 (2012).

158. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* **193,** 693–707 (1987).

159. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics* **18,** 309–317 (1994).

160. Burger, L. & van Nimwegen, E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* **4,** 165 (2008).

161. Skerker, J. M. *et al.* Rewiring the specificity of two-component signal transduction systems. *Cell* **133,** 1043–1054 (2008).

162. Baldassi, C. *et al.* Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PLoS ONE* **9,** e92721 (2014).

163. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* **108,** E1293–301 (2011).

164. Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE* **6,** e28766 (2011).

165. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28,** 184–190 (2012).

166. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* **110,** 15674–15679 (2013).

167. Hopf, T. A. *et al.* Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* (2012). doi:10.1016/j.cell.2012.04.012

168. Nugent, T. & Jones, D. T. Membrane protein structural bioinformatics. *J Struct Biol* **179,** 327–337 (2012).

169. Rajagopala, S. V. *et al.* The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotechnol* **32,** 285–290 (2014).

170. UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* **41,** D43–7 (2013).

171. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42,** D222–30 (2014).

172. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39,** W29–W37 (2011).

173. Rose, P. W. *et al.* The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res* **41,** D475–82 (2013).

174. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11,** 431 (2010).

175. Pakseresht, N. *et al.* Assembly information services in the European Nucleotide Archive. *Nucleic Acids Res* **42,** D38–43 (2014).

176. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87,** 012707 (2013).

177. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics* **77,** 778–795 (2009).

178. Cingolani, G. & Duncan, T. M. Structure of the ATP synthase catalytic complex (F(1)) from Escherichia coli in an autoinhibited conformation. *Nature Publishing Group* **18,** 701–707 (2011).

179. Jacob, F. *The birth of the operon. Science* **332,** 767–767 (American Association for the Advancement of Science, 2011).

180. Aurell, E. & Ekeberg, M. Inverse Ising inference using all the data. *Physical review letters* **108,** 090201 (2012).

181. Hvorup, R. N. *et al.* Asymmetry in the structure of the ABC transporter-binding protein complex BtuCD-BtuF. *Science* **317,** 1387–1390 (2007).

182. Johnson, E., Nguyen, P. T., Yeates, T. O. & Rees, D. C. Inward facing conformations of the MetNI methionine ABC transporter: Implications for the mechanism of transinhibition. *Protein Sci* **21,** 84–96 (2012).

183. Locher, K. P., Lee, A. T. & Rees, D. C. The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science* **296,** 1091–1098 (2002).

184. Kadaba, N. S., Kaiser, J. T., Johnson, E., Lee, A. & Rees, D. C. The high-affinity E. coli methionine ABC transporter: structure and allosteric regulation. *Science* **321,** 250–253 (2008).

185. Rodgers, A. J. & Wilce, M. C. Structure of the gamma-epsilon complex of ATP synthase. *Nature Structural & Molecular Biology* **7,** 1051–1054 (2000).

186. Uhlin, U., Cox, G. B. & Guss, J. M. Crystal structure of the epsilon subunit of the proton-translocating ATP synthase from Escherichia coli. *Structure* **5,** 1219–1230 (1997).

187. Loris, R., Tielker, D., Jaeger, K.-E. & Wyns, L. Structural basis of carbohydrate recognition by the lectin LecB from Pseudomonas aeruginosa. *J Mol Biol* **331,** 861–870 (2003).

188. Stehle, T. & Harrison, S. C. High-resolution structure of a polyomavirus VP1-oligosaccharide complex: implications for assembly and receptor binding. *EMBO J.* **16,** 5139–5148 (1997).

189. Walker, J. E. The ATP synthase: the understood, the uncertain and the unknown. *Biochem Soc Trans* **41,** 1–16 (2013).

190. Baker, L. A., Watt, I. N., Runswick, M. J., Walker, J. E. & Rubinstein, J. L. Arrangement of subunits in intact mammalian mitochondrial ATP synthase determined by cryo-EM. *Proc Natl Acad Sci USA* **109,** 11675–11680 (2012).

191. Long, J. C., DeLeon-Rangel, J. & Vik, S. B. Characterization of the first cytoplasmic loop of subunit a of the Escherichia coli ATP synthase by surface labeling, cross-linking, and mutagenesis. *J Biol Chem* **277,** 27288–27293 (2002).

192. DeLeon-Rangel, J., Zhang, D. & Vik, S. B. The role of transmembrane span 2 in the structure and function of subunit a of the ATP synthase from Escherichia coli. *Arch. Biochem. Biophys.* **418,** 55–62 (2003).

193. Fillingame, R. H. & Steed, P. R. Half channels mediating H(+) transport and the mechanism of gating in the Fo sector of Escherichia coli F1Fo ATP synthase. *Biochim Biophys Acta* **1837,** 1063–1068 (2014).

194. DeLeon-Rangel, J., Ishmukhametov, R. R., Jiang, W., Fillingame, R. H. & Vik, S. B. Interactions between subunits a and b in the rotary ATP synthase as determined by cross-linking. *FEBS Lett* **587,** 892–897 (2013).

195. Dmitriev, O. Y., Jones, P. C. & Fillingame, R. H. Structure of the subunit c oligomer in the F1Fo ATP synthase: model derived from solution structure of the monomer and cross-linking in the native enzyme. *Proc Natl Acad Sci USA* **96,** 7785–7790 (1999).

196. Caviston, T. L., Ketchum, C. J., Sorgen, P. L., Nakamoto, R. K. & Cain, B. D. Identification of an uncoupling mutation affecting the b subunit of F1F0 ATP synthase in Escherichia coli. *FEBS Lett* **429,** 201–206 (1998).

197. Alber, F. *et al.* Determining the architectures of macromolecular assemblies. *Nature* **450,** 683–694 (2007).

198. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys J* **100,** L47–9 (2011).

199. Kozakov, D., Clodfelter, K., Vajda, S. & Camacho, C. Optimal clustering for detecting near-native conformations in protein docking. *Biophys J* **89,** 867–875 (2005).

200. Moult, J. *et al.* Critical assessment of methods of protein structure prediction-Round VII. *Proteins: Structure, Function, and Bioinformatics* **69,** 3–9 (2007).

201. Reva, B. A., Finkelstein, A. V. & Skolnick, J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 å? *Folding and Design* **3,** 141–147 (1998).

202. Wallin, S., Farwer, J. & Bastolla, U. Testing similarity measures with continuous and discrete protein models. *Proteins: Structure, Function, and Bioinformatics* **50,** 144–157 (2003).

203. Cossio, P., Laio, A. & Pietrucci, F. Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? *Phys. Chem. Chem. Phys.* **13,** 10421–10425 (2011).

204. Mariani, V., Kiefer, F., Schmidt, T., Haas, J. & Schwede, T. Assessment of template based protein structure predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics* n/a–n/a (2011). doi:10.1002/prot.23177

205. Prinzie, A. & Van den Poel, D. Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM. *Decision Support Systems* **42,** 508–526 (2006).

206. Wang, G. *et al.* Solution structure of the phosphoryl transfer complex between the signal transducing proteins HPr and IIA(glucose) of the Escherichia coli phosphoenolpyruvate:sugar phosphotransferase system. *EMBO J.* **19,** 5635–5649 (2000).

207. Buckle, A. M., Schreiber, G. & Fersht, A. R. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. *Biochemistry* **33,** 8878–8889 (1994).

208. Rey, F. A., Heinz, F. X., Mandl, C., Kunz, C. & Harrison, S. C. The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. *Nature* **375,** 291–298 (1995).

209. Horton, J. R. & Cheng, X. PvuII endonuclease contains two calcium ions in active sites. *J Mol Biol* **300,** 1049–1056 (2000).

210. Daura, X. *et al.* Peptide folding: when simulation meets experiment. *Angewandte Chemie International Edition* **38,** 236–240 (1999).

211. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. From words to literature in structural proteomics. *Nature* **422,** 216–225 (2003).

212. Voorhees, R. M., Fernández, I. S., Scheres, S. H. W. & Hegde, R. S. Structure of the Mammalian ribosome-sec61 complex to 3.4 å resolution. *Cell* **157,** 1632–1643 (2014).

213. Berman, H. M., Kleywegt, G. J., Nakamura, H. & Markley, J. L. The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure* **20,** 391–396 (2012).

214. Anger, A. M. *et al.* Structures of the human and Drosophila 80S ribosome. *Nature* **497,** 80–85 (2013).

215. Galej, W. P., Nguyen, T. H. D., Newman, A. J. & Nagai, K. Structural studies of the spliceosome: zooming into the heart of the machine. *Curr Opin Struct Biol* **25C,** 57–66 (2014).

216. Bateman, A. & Valencia, A. Structural genomics meets computational biology. *Bioinformatics* **22,** 2319 (2006).

217. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) - round X. *Proteins: Structure, Function, and Bioinformatics* (2013). doi:10.1002/prot.24452

218. D'Angelo, M. & Hetzer, M. Structure, dynamics and function of nuclear pore complexes. *Trends in cell biology* **18,** 456–466 (2008).

219. Lasker, K. *et al.* Integrative structure modeling of macromolecular assemblies from proteomics data. *Molecular & Cellular Proteomics* (2010).

220. Nilges, M. & O'Donoghue, S. I. Ambiguous NOEs and automated NOE assignment. *Progress in Nuclear Magnetic Resonance Spectroscopy* **32,** 107–139 (1998).

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

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

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

224. Snijder, J. *et al.* Insight into cyanobacterial circadian timing from structural details of the KaiB-KaiC interaction. *Proc Natl Acad Sci USA* **111,** 1379–1384 (2014).

225. Thomson, J. J. XXIV. On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *Philosophical Magazine* **7,** 237–265 (1904).

226. Erber, T. & Hockney, G. M. Equilibrium configurations of N equal charges on a sphere. *J. Phys. A* **24,** L1369–L1378. 12 p (1991).

227. Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* **104,** 59–107 (1976).

228. Sansom, M. S. P., Scott, K. A. & Bond, P. J. Coarse-grained simulation: a high-throughput computational approach to membrane proteins. *Biochem Soc Trans* **36,** 27–32 (2008).

229. Baaden, M. & Marrink, S. J. Coarse-grain modelling of protein-protein interactions. *Curr Opin Struct Biol* **23,** 878–886 (2013).

230. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & de Vries, A. H. The MARTINI force field: coarse

grained model for biomolecular simulations. *J Phys Chem B* **111,** 7812–7824 (2007).

231. Monticelli, L. *et al.* The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput* **4,** 819–834 (2008).

232. de Jong, D. H. *et al.* Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput* **9,** 687–697 (2013).

233. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).

234. Schüttelkopf, A. W. & van Aalten, D. M. F. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr* **60,** 1355–1363 (2004).

235. Sousa da Silva, A. W. & Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res Notes* **5,** 367 (2012).

236. Freddolino, P. L., Liu, F., Gruebele, M. & Schulten, K. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys J* **94,** L75–7 (2008).

237. Shaw, D. E. *et al.* Atomic-level characterization of the structural dynamics of proteins. *Science* **330,** 341–346 (2010).

238. Kastritis, P. L. & Bonvin, A. M. J. J. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.* **9,** 2216–2225 (2010).

239. Zhang, J., Zheng, F. & Grigoryan, G. Design and designability of protein-based assemblies. *Curr Opin Struct Biol* **27C,** 79–86 (2014).

240. Cooper, S. *et al.* Predicting protein structures with a multiplayer online game. *Nature* **466,** 756–760 (2010).

241. Levieux, G. *et al.* Udock, the interactive docking entertainment system. *Faraday Discuss.* (2014). doi:10.1039/C3FD00147D

242. Baker, D. & Group, F. C. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & …* (2011).

243. Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S. & Khatib, F. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature* (2012).

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

245. Heijne, von, G. The membrane protein universe: what's out there and why bother? *J. Intern. Med.* **261,** 543–557 (2007).

246. Dancea, F., Kami, K. & Overduin, M. Lipid interaction networks of peripheral membrane proteins revealed by data-driven micelle docking. *Biophys J* **94,** 515–524 (2008).

247. Lindahl, E. & Sansom, M. S. P. Membrane proteins: molecular dynamics simulations. *Curr Opin Struct Biol* **18,** 425–431 (2008).

248. Feig, M. Implicit membrane models for membrane protein simulation. *Methods Mol Biol* **443,** 181–196 (2008).

249. Hong, M., Zhang, Y. & Hu, F. Membrane Protein Structure and Dynamics from NMR Spectroscopy. *Annu Rev Phys Chem* **63,** 1–24 (2012).

250. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat Rev Drug Discov* **5,** 993–996 (2006).

251. White, S. Biophysical dissection of membrane proteins. *Nature* **459,** 344–346 (2009).

252. Kelm, S., Shi, J. & Deane, C. M. iMembrane: homology-based membrane-insertion of proteins. *Bioinformatics* **25,** 1086–1088 (2009).

253. Kelm, S., Shi, J. & Deane, C. M. MEDELLER: homology-based coordinate generation for membrane proteins. *Bioinformatics* **26,** 2833–2840 (2010).

Summary

Protein interactions are critical to cellular homeostasis. This sentence, so often repeated in abstracts and introductions, is an understatement to the importance of protein molecules and their interactions. In a bacterial cell, which is smaller and simpler than our own, proteins make up for more than half of its dry weight. In comparison, DNA, the molecule of life, accounts for a much smaller percentage: ~3%. The justification for such abundance is simple: protein molecules are the building blocks, the builders, and the maintainers of the cell. They constitute the majority of the cytoskeleton, regulate protein production via ribosomal synthesis, oversee their own folding process via chaperones, and when the time comes, process their own degradation via the proteasome. More over, proteins help maintain the structural and chemical integrity of other molecules, such as DNA. For example, when ultraviolet radiation acts on the genetic material in the nucleus, some proteins, such as XPF and ERCC1, repair the damages, or mutations, at the molecular level while others synthesize pigments, such as melanin, which absorb further radiation. The success of all these processes, and in fact of many others, hinges on the interactions between proteins and of those between proteins and other macromolecules. The degree of accuracy in these interactions is astonishing and dwarfs many engineering feats of humankind, particularly given the estimated number of interactions occurring in each cell: more than one hundred thousand identified unique interactions in a human cell, many copies of each occurring simultaneously at any given moment in each cell of our bodies. Indeed, and more remarkably, many of these interactions exist to compensate for when others fail to function adequately, showing that despite that preciseness, there are built-in fail-safe mechanisms to rescue the cell from a state of disarray. Inevitably, and in line with Murphy's Law, eventually the backup fails, leading to conditions we associate with disease.

For obvious reasons, the scientific community has been ardently devoted to the understanding of protein interactions. Given a collection of possibly interacting proteins, not only is it important to discover which ones do interact and which process they possibly regulate or carry out, but also *how* exactly these interactions take place. Frequently, the *how* is the last question being answered, because of technical and methodological difficulties. Elucidating the details of a single interaction requires the atomic structure determination of the partner proteins and their complex, by means of X-ray crystallography and/or other experimental biophysical techniques, which are never trivial to execute despite decades of developments. Fortunately, other technologies emerged to provide support to experiments, and in some cases act as the only viable alternative, such as computer-assisted modeling based on physico-chemical principles. Parallel to the explosive developments in computer science of the last decades, the application of these computational methods to protein-protein interactions underwent a significant evolution since their inception in the early 1970s. The field of Computational Structural Biology stands nowadays as an independent and mature branch of science, whose tools are valuable assets for the overall biological community.

Yet, there are unmet challenges in the field. With experimental researchers becoming more ambitious, spurred by the technology developments in structure determination techniques, computational methods must continuously evolve and adapt. Of particular interest to this thesis are the problems of dealing with large molecular machines and assemblies, and specifically those cases where there is limited information, experimental or predicted, to obtain a three-dimensional model.

In the **General Introduction**, I offer a chronological overview spanning the last two centuries, from the scientific discoveries that led to the establishment of proteins as key players in cellular metabolism, to the establishment and development of structural biology

methods that provide the much needed atomic-resolution information. This brief historical account contextualizes and highlights, to the general reader, the importance of the creation and development of Computational Structural biology.

Following this is a more in-depth introduction to the methods used by computational structural biologists in the study of protein-protein interactions. In **Chapter 1**, I first define computational docking, the problem of predicting the structure of a (protein) complex from its free molecular components. I then proceed to summarize the different approaches used in protein-protein docking, distinguishing those that make use of external information derived from experiments or predictions – integrative or data-driven modeling – from those that rely solely on physico-chemical principles – *ab initio* docking. Further, I emphasize the advantages of integrative approaches and show their success in the community-wide blind experiment CAPRI. In addition to the methodology details, I list and explain the different types and sources of information that current integrative modeling approaches support. Towards the end of the chapter, I briefly venture into the field of protein interface prediction, which will be important to understand the impact of the developments laid out in **Chapter 3**. For the sake of completeness, I conclude this chapter with a summary of alternative methods to docking and their advantages and disadvantages.

Chapter 2 explores the practice of homology modeling in the context of docking, for cases where (at least) one of the interacting partners does not have an experimentally determined structure. This is becoming more common as the number of protein structures grows, together with the availability and ease-of-use of web-based homology modeling tools. High identity homologues of the partners will be structurally closer to the reference structure and thus can produce better models after docking. However, how much different is a docked model after starting from a more distant homologue? Is there a measure to predict the achievable quality of the docked model using the starting homology models? After collecting a dataset of protein-protein complexes and searching for homologues, I assess several different metrics, some as simple as the sequence identity between target and template, and others created specifically for this purpose, such as the TVSMod_RMSD score. I show that the sequence identity is the simplest and best indicator of success, as it correlates best with the structural similarity of the docked models to the crystal structure of the complex. Finally, I assess the impact of the quality of the data on the final docking models. Each of the protein complexes in the dataset was once an unknown target presented to our research group as part of the CAPRI experiment. For each, we had to search for clues (i.e., from published literature, predictions, related complexes) on the region that defined the interface. I compare the results of using both this set of data and another derived from the crystal structure of the complex, i.e., perfect information, in docking with the different homologues, and conclude that the quality of the starting models is secondary to the quality of the information used to guide the modeling. Combined, the findings in this chapter shed light on two important steps of a docking calculation and suggest that instead of focusing on the quality of the homology model, researchers looking to build a model of their interaction should divert their efforts to either producing or finding better interface information.

Sometimes, however, the particular interaction under study is not easy to characterize experimentally, either because of the very weak affinity of the interacting partners, or because of very particular biophysical requisites (i.e., ligands, salt concentrations, solubility, membrane environment). In such cases, the only option to produce clues on possible interfaces comes also from computational methods: interface prediction. **Chapter 3** introduces a recently developed class of methods based on the theory of protein coevolution. These

methods assume that amino acids that are neighbors in the three-dimensional structure exert an evolutionary pressure on one another in order to keep their spatial relationship, as protein structure defines protein function. The same rationale applies to amino acids bridging the interaction of two proteins. Together with my collaborators, I show that this assumption is valid and that indeed it is possible to derive spatial information between interacting proteins, using only information contained in their evolutionary history. By using the results of these predictions as distance restraints in HADDOCK, it was possible to obtain near-native models for several bacterial protein complexes. These results indicate a possible high-accuracy method to predict interfaces between proteins, and suggest several improvements that ought materialize in order to pave the way for genome-wide applications.

Finally, the last two chapters are dedicated to advances in computational methods to predict the structure of large molecular machines, such as the ribosome. Until recently, the sheer size and complexity of these molecular assemblies prevented their dissection through both experimental and computational methods. Advances in molecular biology and structural biology pushed the limits of conventional structure determination techniques, while the development of other methods such as cryo-electron microscopy, which has recently achieved resolutions nearing those of X-ray crystallography and small-angle scattering provided new opportunities to study the structure of large biological systems. Accordingly, computational structural biology methods need to adapt to these challenging systems.

First, in **Chapter 4**, I describe a new clustering algorithm designed specifically for models of molecular interactions. Modeling methods generate a large number of models to counterbalance the inaccuracy of their sampling and scoring protocols. Afterwards, it is common to use clustering methods to identify a large population of structurally similar low-energy models and consider it as representative of the native model. This hypothesis derives from observations in both molecular dynamics and protein docking studies. Another goal of clustering is to eliminate rare low-scoring models that might originate from fortuitous errors during sampling or scoring. Traditional clustering algorithms rely on the root mean square deviation of atomic coordinates to infer structural similarity, a measure that requires an *a priori* spatial superimposition or fitting step. Consequently, performing this fitting operation for a large number of models is computationally expensive, particularly in the case of flexible docking methods, such as HADDOCK. Moreover, large assemblies composed of multiple copies of the same molecule often display conformational symmetry. When building such a model, the order of the individual partners is not always respected (i.e., ABC vs. ACB), which leads to additional fitting calculations, specifically *(n-1)!* calculations for *n* molecules. The algorithm I describe in **Chapter 4** is computationally more efficient since it relies on a contact-based measure that implicitly describes structural similarity and does not require any *a priori* fitting, and is able to bypass the symmetry-imposed extra calculations with a simple data manipulation. It is thus better suited for large-scale docking studies, such as whole-genome docking, to molecular systems composed of several subunits, and to symmetrical assemblies, therefore presenting itself as an attractive method for docking approaches.

Then, in **Chapter 5**, I present a major update to the integrative modeling software developed by the group of Alexandre Bonvin at Utrecht University, HADDOCK. Originally developed in 2003, HADDOCK pioneered the use of experimental data in docking, by making use of nuclear magnetic resonance spectroscopy and mutagenesis data to drive the calculations. In the following years, the group added other sources of information to HADDOCK's repertoire and extended the software to be able to dock up to six molecules

simultaneously. However, in light of the size and complexity of recent structural biology studies (e.g., the ribosome, the spliceosome, 20s proteasome, etc.), other developments were in order to keep HADDOCK *on par* with experimental research. I describe two significant changes made to the HADDOCK protocol. First, I detail the design and implementation of an initial placement algorithm for an arbitrary number of molecules, which effectively lifts the six-molecule limitation. I demonstrate the impact of this algorithm on the quality of the docking models for three protein complexes when compared to the original implementation. I also investigate the parameters used in the initial rigid-body conformational search stage, originally designed for two-body systems and I show that the current settings are too strict to allow optimal results for larger systems. Finally, I demonstrate the successful implementation of the MARTINI force field in HADDOCK and highlight the advantages of using a coarse-grained representation, which reduces the number of particles, particularly regarding computational efficiency and conformational sampling of large and complex systems. These developments will allow HADDOCK to model the ambitious biological systems currently targeted by experimental researchers, and thus open a venue for the exploration of uncharted territories of the structural interactome.

Samenvatting

Interacties tussen eiwitten zijn van cruciaal belang voor de homeostase van een cel. Deze zin, die zeer vaak voorkomt in samenvattingen of introducties, geeft slechts zacht gezegd aan wat het belang is van eiwitmoleculen en hun interacties. In bacteriële cellen, die over het algemeen kleiner zijn dan onze eigen cellen, bepalen eiwitten meer dan de helft van het droge gewicht. Ter vergelijking, DNA, het molecuul dat onze genetische informatie bevat, zorgt slechts voor ~3% van het droge gewicht. De reden voor deze hoge dichtheid aan eiwitten is simpel: eiwit moleculen vormen de bouwblokken, de bouwers, als ook de opzichters van de cel. Ze vormen een groot deel van het cytoskelet van de cel, reguleren de productie van eiwitten door middel van het ribosoom, houden toezicht op hun eigen vouwingsproces via chaperonne-eiwitten, en wanneer de tijd rijp is, verzorgen ze hun eigen afbraak gebruikmakend van het proteasoom. Daarnaast helpen eiwitten bij het intact houden van de structuur en het chemische karakter van andere moleculen zoals DNA. Als bijvoorbeeld ultraviolette straling het genetische materiaal in de kern aantast, dan zullen specifieke eiwitten zoals XPF en ERCC1 de schade herstellen of de aangebrachte mutaties terugdraaien, terwijl andere eiwitten pigmenten, zoals melanine, produceren om de straling te absorberen. Het slagen van al deze processen en nog vele andere hangt af van interacties tussen eiwitten en van eiwitten met andere macromoleculen. Het is verbazingwekkend hoe nauwkeurig deze interacties zijn en menig technische prestatie van mensenhand valt erbij in het niets. Zeker als je bedenkt hoeveel interacties er ongeveer plaatsvinden in iedere cel: in een menselijke cel meer dan honderdduizend unieke interacties, waarvan een groot deel tegelijkertijd in iedere cel van ons lichaam plaatsvindt. Opvallend is dan ook dat de rol van veel van deze interacties het compenseren voor problemen is, in het geval dat andere interacties niet goed hun functie uitvoeren. Dit laat zien dat, ondanks de hoge precisie, er ingebouwde veiligheidsmechanismes zijn om te voorkomen dat het een totale wanorde wordt in de cel. Helaas zal uiteindelijk, zoals de wet van Murphy zegt, ook de back-up falen, wat vervolgens zorgt voor een situatie die we met een ziektebeeld associëren.

Het ligt dus voor de hand dat de wetenschapsgemeenschap met verwoede toewijding eiwitinteracties probeert te begrijpen. Voor een set van eiwitten die mogelijkerwijs interacteren is het namelijk niet alleen belangrijk te ontdekken welke interacties daadwerkelijk plaatsvinden en welk proces daarmee gereguleerd of uitgevoerd wordt, maar ook te begrijpen *hoe* deze interacties precies plaatsvinden. Vaak is *hoe* echter de laatste vraag die beantwoord wordt, voornamelijk door problemen van technische en methodische aard. Om de details van één specifieke interactie te begrijpen, moet de atomaire structuur van beide partnereiwitten en ook hun complex met behulp van röntgendiffractie en/of andere biofysische technieken bepaald worden, wat ondanks de vele ontwikkelingen van de afgelopen decennia nog altijd niet triviaal is. Gelukkig zijn er technologieën ontwikkeld om de experimentele technieken bij te staan, die soms zelfs het enige alternatief zijn, zoals het computergestuurd modeleren gebaseerd op fysisch-chemische principes. Naast de enorme ontwikkelingen van de informatica in de laatste decennia, is de toepassing van deze computationele methoden voor de studie van eiwit-eiwit interacties, na hun introductie aan het begin van de zeventiger jaren, aanzienlijk geëvolueerd. De Computationele Structuur Biologie is nu een zelfstandig en volgroeid onderzoeksveld en de ontwikkelde programma's zijn waardevol voor de gehele biologische community.

Toch zijn er nog genoeg uitdagingen voor het onderzoeksveld. Experimentele onderzoekers worden, door de continue ontwikkelingen van de structuur bepalingstechnieken, steeds ambitieuzer, waardoor computationele methoden moeten blijven veranderen en ontwikkelen. In dit proefschrift komen voornamelijk problemen aan bod die te maken

hebben met grote moleculaire machines en assemblages, en gevallen waar er weinig experimentele of voorspelde informatie is voor het verkrijgen van een driedimensionaal model.

In **Hoofdstuk 1** geef ik een chronologisch overzicht van de wetenschappelijke uitvindingen van de laatste twee decennia, die ervoor gezorgd hebben dat eiwitten konden worden geïdentificeerd als de sleutelfiguren van het cellulaire metabolisme, tot aan de ontwikkeling van methoden in de structuur biologie die de zo belangrijke atomaire resolutie verschaffen. Dit korte historische overzicht legt aan de algemene lezer het belang van het onderzoeksveld van de Computationele Structuur Biologie uit.

Hierna volgt een meer diepgaande introductie van de methoden die door computationele structuur biologen worden gebruikt voor het bestuderen van eiwit-eiwit interacties. In **Hoofdstuk 2** definieer ik eerst wat computationeel dokken is, namelijk het voorspellen van een (eiwit)complex gebruikmakend van de vrije moleculaire componenten. Vervolgens beschrijf ik de verschillende methoden die gebruikt worden bij het eiwit-eiwit dokken, waarbij onderscheid gemaakt kan worden tussen methoden die gebruikmaken van experimentele of voorspelde informatie – het geïntegreerd of data-aangedreven dokken – en methoden die uitsluitend fysisch-chemische principes gebruiken – het *ab initio* dokken. Verder benadruk ik de voordelen van methoden gebaseerd op het geïntegreerd dokken en laat ik zien hoe goed deze presteren in het CAPRI experiment. Naast de details van de methoden, geef ik een overzicht van de verschillende type informatie die de huidige geïntegreerde methoden kunnen gebruiken en ik licht deze ook toe. Aan het einde van het hoofdstuk maak ik een klein uitstapje in het onderzoeksveld van de eiwitraakvlak voorspellingen. Dit, om later de impact van de ontwikkelingen beschreven in **Hoofdstuk 4** in te kunnen schatten. Ter volledigheid sluit ik het hoofdstuk af met een overzicht van andere alternatieven naast dokken, inclusief hun voor- en nadelen.

In **Hoofdstuk 3** wordt het modeleren op basis van homologie uitgewerkt, wat gebruikt wordt in het geval dat van (minstens) één van de partnereiwitten de experimentele structuur niet bepaald is. Dit wordt steeds vaker gebruikt, onder andere door de groei van het aantal eiwitstructuren en de toenemende toegankelijkheid en gebruiksvriendelijkheid van programma's op het web voor het modeleren op basis van homologie. Homologen van partners met veel identieke residuen zullen meer op de referentie structuur lijken en vervolgens dus ook voor een beter resultaat zorgen na het dokken. Maar hoe anders is een gedokt model wanneer een minder identieke homoloog gebruikt wordt? En is er een manier om te voorspellen welke kwaliteit het gedokte model kan halen wanneer gebruik gemaakt wordt van homologie modellen? Na een dataset van eiwit-eiwitcomplexen te hebben samengesteld en de bijbehorende homologen te hebben verzameld, heb ik verschillende parameters getoetst, zoals de mate van identiteit tussen de sequenties van het doelwit en het sjablooneiwit, maar ook andere speciaal gecreëerde parameters zoals de TVSMod_RMSD score. Ik laat zien dat de mate van identiteit de simpelste en beste indicator voor succes is, omdat deze het beste correleert met de mate waarop de gedokte modellen gelijkenis vertonen met de kristalstructuur van het complex. Alle eiwitcomplexen in de dataset waren eerder ook een casus in het CAPRI experiment, waaraan onze onderzoeksgroep sinds jaren deelneemt. In alle gevallen moesten we informatie (uit de literatuur, van voorspellingen, op basis van gerelateerde complexen etc.) verzamelen over de regio in het eiwit die als raakvlak gedefinieerd moest worden. Ik vergelijk de resultaten van het dokken met homologen, verkregen op basis van deze informatie, met het gebruik van "perfecte" informatie, welke gebaseerd is op de kristalstructuur. Hieruit volgt de conclusie dat de kwaliteit van de startmodellen secundair is ten opzichte van de informatie die gebruikt

wordt om het dokken in goede banen te leiden. De bevindingen in dit hoofdstuk belichten twee belangrijke etappes van een dok-berekening en bevelen onderzoekers, die een model van een bepaalde interactie willen berekenen, aan hun energie te steken in het genereren van of zoeken naar betere informatie over het raakvlak, in plaats van tijd te spenderen aan de kwaliteit van het homologie model.

In sommige gevallen is een bepaalde interactie echter niet gemakkelijk experimenteel te karakteriseren, bijvoorbeeld door een zeer zwakke affiniteit van de bindingspartners, of omdat specifieke condities (zoals liganden, zout concentratie, oplosbaarheid of membraan omgeving) aan de orde zijn. In deze gevallen is de enige optie, om toch informatie te verkrijgen over het raakvlak, ook hier een computergestuurde methode te gebruiken. In **Hoofdstuk 4** worden een aantal recent ontwikkelde methoden besproken, die gebaseerd zijn op de theorie van de co-evolutie. Deze methoden gaan ervan uit dat aminozuren, die qua afstand in de 3D-structuur elkaars buren zijn, een evolutionaire druk op elkaar uitoefenen, die ervoor zorgt dat hun afstand ten opzichte van elkaar in stand blijft. Het idee hierachter is dat de structuur van een eiwit de functie van een eiwit bepaald. Dezelfde denkwijze kan worden toegepast op aminozuren die een koppel vormen tussen twee interacterende eiwitten. In samenwerking met andere onderzoekers, heb ik laten zien dat deze aanname juist is en dat het inderdaad mogelijk is afstandsinformatie te verkrijgen, uitsluitend gebruikmakend van de evolutionaire geschiedenis. Door deze informatie als beteugeling (*restraints)* in HADDOCK te gebruiken, was het mogelijk om modellen van verschillende bacteriële eiwitcomplexen te verkrijgen die zeer sterk op de native structuur lijken. Dit geeft aan dat deze methode mogelijkerwijs zeer accuraat is in het voorspellen van het raakvlak tussen twee eiwitten en in de toekomst, na het uitwerken van een aantal voorgestelde verbeteringen, wellicht gebruikt kan worden voor genoom-brede studies.

De laatste twee hoofdstukken zijn gereserveerd voor ontwikkelingen in de computationele methoden voor het voorspellen van de structuur van grote moleculaire machines, zoals het ribosoom. Tot voor kort waren de grootte en complexe samenstelling van deze moleculaire assemblages problematisch bij het bestuderen van hun structuur met behulp van experimentele of computationele methoden. Maar ontwikkelingen in zowel de moleculaire biologie als ook de structuur biologie hebben ervoor gezorgd dat de grenzen van de haalbaarheid voor de conventionele structuurbepalingstechnieken zijn verschoven. Daarnaast zorgt de ontwikkeling van andere methoden, zoals cryo-elektronenmicroscopie, waarbij recentelijk resoluties zijn gehaald die in de buurt van röntgendiffractie en kleine-hoekverstrooiing komen, voor nieuwe mogelijkheden om de structuur van dit soort grote biologische systemen te bestuderen. Natuurlijkerwijs moeten de computationele methoden ook aangepast worden aan deze uitdagende systemen.

Eerst beschrijf ik in **Hoofdstuk 5** een nieuwe algoritme voor het clusteren van modellen, wat specifiek ontwikkeld is voor modellen van moleculaire interacties. Methoden voor het berekenen van modellen genereren vaak een zeer groot aantal modellen om tegenwicht te geven aan de over het algemeen lage accuraatheid van de bemonstering- en scoringsprotocollen. Na een berekening is het de gewoonte om een clustermethode te gebruiken om een grote populatie van gelijksoortige lage energie structuren te kunnen identificeren, om deze te kunnen overwegen als goede representatie van de native structuur. Deze hypothese komt voort uit observaties van zowel de moleculaire dynamica als studies met eiwitdokken. Een ander doel van het clusteren is om zeldzame modellen met een lage score, die mogelijkerwijs voortkomen uit incidentele fouten tijdens het bemonsteren of scoren, te elimineren. Traditionele cluster algoritmes gebruiken de kwadratisch gemiddelde afwijking (RMSD) van de

atomaire coördinaten om de structurele gelijkheid te bepalen, waarbij een *a priori* overlap of passingsstap nodig is. Derhalve kost deze stap veel computerkracht wanneer een groot aantal modellen gepast moeten worden, wat typisch het geval is bij flexibele dokmethoden zoals HADDOCK. Daarnaast vertonen grote assemblages die bestaan uit meerdere kopieën van hetzelfde molecuul vaak conformationele symmetrie. Wanneer zo'n soort model opgebouwd wordt is de volgorde van de individuele partners niet altijd hetzelfde (b.v. ABC vs. ACB) wat zorgt voor additionele passingsberekeningen, om precies te zijn *(n-1)!* berekeningen voor *n* moleculen. Het algoritme dat ik in **Hoofdstuk 5** beschrijf is computationeel veel efficiënter, omdat het gebaseerd is op de aanwezigheid van contacten, wat op een impliciete manier bepaald of structuren gelijk zijn waardoor er geen *a priori* passing meer nodig is en de extra berekeningen in het geval van symmetrie omzeild worden. Dit algoritme is dus beter geschikt voor het op grote schaal dokken, zoals het dokken van het gehele genoom, moleculaire systemen die opgebouwd zijn uit vele moleculen en symmetrische assemblages, waardoor het een aantrekkelijke methode is.

Vervolgens presenteer ik in **Hoofdstuk 6** een belangrijke *update* van de HADDOCK *software* voor het geïntegreerd modeleren, die ontwikkeld is door de onderzoeksgroep van Alexandre Bonvin aan de Universiteit Utrecht. Toen HADDOCK in 2003 ontwikkeld werd was het baanbrekend dat experimentele data, verkregen met behulp van kernspinresonantie (NMR) spectroscopie en mutagenese, bij het dokken gebruikt werden om de berekening te sturen. In de jaren die daar op volgden, werden nieuwe bronnen van informatie aan het HADDOCK repertoire toegevoegd en werd de *software* uitgebreid om tot aan zes moleculen tegelijk te kunnen dokken. Maar in het licht van de grootte en complexiteit van de huidige structuur biologie studies (het ribosome, het spliceosoom, de 20S proteasome, etc.) was het nodig andere facetten binnen HADDOCK te ontwikkelen om het programma in lijn te houden met het experimentele onderzoek. Ik beschrijf twee belangrijke aanpassingen van het HADDOCK protocol. Ten eerste ga ik in op het design en de implementatie van een initieel plaatsingsalgoritme voor een arbitrair aantal moleculen, waardoor de limiet van zes moleculen effectief opgeheven is. Ik laat voor drie eiwitcomplexen zien wat de impact van dit algoritme op de kwaliteit van de gedokte modellen is in vergelijking met het originele protocol. Ik onderzoek ook de parameters die gebruikt worden in de initiële star-lichaamfase, die in eerste instantie ontwikkeld was voor twee-lichaamsystemen en ik laat zien dat de huidige parameters te strikt zijn gezet om optimale resultaten te verkrijgen voor grotere systemen. Als laatste beschrijf ik de succesvolle implementatie van het MARTINI krachtenveld in HADDOCK en benoem ik de voordelen van een grof korellige representatie die het aantal deeltjes verkleind, met name wanneer gekeken wordt naar computationele efficiëntie en conformationele bemonstering van grote en complexe systemen. Deze ontwikkelingen zorgen ervoor dat HADDOCK de ambitieuze systemen, die op het moment door experimentele onderzoekers onder de loep worden genomen, zal kunnen modeleren en dit opent dus nieuwe wegen voor een ontdekkingsreis langs de nog onbegane gebieden van het structurele interactoom.

Resumo

Não há ano sem que, durante o tradicional jantar de Natal, o meu avô pouse o copo e os talheres, ajeite os óculos, e me faça *aquela* pergunta: "Mas afinal, explica-me lá, o que é que tu fazes no laboratório?". É daqueles momentos em que os créditos de licenciatura, aulas de mestrado e anos de doutoramento de pouco ou nada valem porque, ao fim e ao cabo, se nem aos meus colegas de curso consigo explicar em condições, que esperança devo ter de o fazer ao meu avô? Acho que ainda lhe devo uma boa explicação. Esta não é portanto uma tradução literal do resumo em Inglês (e Holandês), mas uma versão diferente, mais simples, que espero que dê resposta à *tal* pergunta.

Comecemos pelas células. Cada um de nós é feito de células, cerca de 37.200.000.000 (trinta e sete biliões e duzentos mil milhões) diferenciadas nuns duzentos tipos diferentes, cada um com uma função bastante específica. Há células cardíacas, que contraem ritmicamente para fazer circular o sangue, células cerebrais especializadas em produzir e transmitir impulsos eléctricos, ou células intestinais cuja tarefa é absorver certos nutrientes dos alimentos que ingerimos. Apesar das diferenças em tamanho, forma e função, a maioria das células partilha um mesmo elemento base: o genoma, ou, mais especificamente, o ácido desoxirribonucleico (ADN). Neste encontram-se codificadas as instruções para construir uma outra classe especial de moléculas, as proteínas, que são responsáveis por (quase) *todas* as funções celulares.

Uma proteína é uma combinação única de moléculas chamadas aminoácidos (p. ex. glutamato, triptofano) que se dobram e orientam em três dimensões e definem a identidade e função da proteína. Estima-se que o ADN das células humanas tenha informação para codificar onze mil tipos de proteínas diferentes, e que em cada célula humana existam, num determinado momento, várias cópias de cada uma, num total de 8.000.000.000 (oito mil milhões) de proteínas. A função desempenhada por cada proteína varia bastante: umas constituem a infraestrutura base - fundações, estradas, etc. - da célula, enquanto outras mantém a informação contida no ADN, corrigindo erros (mutações) que vão sendo introduzidos ao longo do tempo; algumas são responsáveis por sintetizar outras proteínas, outras ainda por destruir proteínas que, ou já não são necessárias, ou deixaram de conseguir cumprir a sua função.

Curiosamente, a maioria destas funções são levadas a cabo, não pelas proteínas individualmente, mas por grupos de proteínas em conjunto. Cada interacção entre proteínas, e entre estas e outras moléculas, é um evento bastante específico e com um resultado bem definido. Qualquer alteração nestas interacções, seja por agentes exógenos, tais como venenos, ou endógenos, como mutações genéticas, tem frequentemente consequências drásticas para célula e para o organismo em geral. Por exemplo, o oxigénio que respiramos é utilizado na produção de energia na célula, na forma de ATP, através de uma sucessão de interacções entre proteínas e entre proteínas e pequenas moléculas. A inalação de monóxido de carbono, ou a ingestão de cianeto, interrompem uma destas interacções e impedem a progressão do processo de formação de ATP, causando então morte celular. Outro exemplo, infelizmente mais comum, é o caso da proteína p53. Esta proteína, apelidada de "guardiã do genoma", interage com inúmeras outras proteínas para evitar a propagação de mutações no ADN. No caso de mutações fortuitas no gene da própria p53, esta deixa de poder interagir correctamente com as suas parceiras e compromete assim a integridade do material genético. Consequentemente, futuras mutações tornam-se mais frequentes e a sua acumulação pode levar a célula a tornar-se cancerígena.

Dada a importância das interacções entre proteínas para o bem-estar da célula, é natural que os cientistas, em particular biólogos moleculares e bioquímicos, se dediquem a tentar perceber quando, porquê e, principalmente, como é que estas se dão. A indústria farmacêutica também é parte interessada neste problema, em particular dada a importância destas interacções e da sua manipulação para fins terapêuticos. No entanto, o tamanho diminuto das proteínas bem

como a duração das suas interacções tornam tais estudos bastante complicados. A hemoglobina, por exemplo, um complexo proteico formado por quatro proteínas diferentes e cuja função é o transporte de oxigénio no sangue, mede apenas 5 nanómetros de diâmetro. Pondo este número em perspectiva, se a uma garrafa de água fosse do tamanho de Portugal, uma molécula de hemoglobina seria do tamanho de uma garrafa de água. Agora, imagine-se que a cada segundo no nosso corpo morrem meio milhão de células por morte programada, e que cada morte é o resultado de milhões de interacções entre proteínas. Cada interacção tem obrigatoriamente que acontecer então no espaço de poucos nanosegundos.

Como é que é possível então observar e estudar estas interacções? A resposta passa essencialmente por uma técnica laboratorial que permite tirar "fotografias" a proteínas: cristalografia de raios-X. Outra técnica, mais recente e complexa, é a ressonância magnética nuclear, que "escuta" os aminoácidos que compõem as proteínas e, sabendo quais comunicam com quais, consegue deduzir como é que as proteínas interagem. Ambas são invenções magníficas e foram galardoadas com vários prémios Nobel. Porém, dada a morosidade em obter resultados e as despesas associadas com o equipamento, conhecemos apenas a estrutura de cerca de 4.200 das 50.000 interacções entre proteínas da célula humana, apesar de quase cinco décadas de investigações. A este ritmo, precisaremos de uns quantos séculos para elucidar todo o universo de interacções nas células humanas. Depois dessas ainda faltarão as das células bacterianas, assim como aquelas entre as proteínas virais e as nossas, etc.

Uma alternativa a estas técnicas experimentais é o uso de computadores. Se os computadores actuais são poderosos o suficiente para substituir cenários e actores em vários filmes de Hollywood, porque não utilizá-los para fins mais nobres? Desde o início dos anos setenta que vários grupos de investigação têm vindo a programar maneiras de modelar e simular proteínas e, em especial, as suas interacções, usando computadores - o chamado *docking*. Conhecendo a forma de duas proteínas e sabendo que estas interagem, é possível escrever um programa informático que tenta adivinhar como é que a interacção se dá na realidade, produzindo aquilo a que chamamos um modelo. É como aquele puzzle para bebés em que se tem que encaixar uma forma num determinado lugar: a estrela não encaixa no lugar do cubo e vice-versa. Entre proteínas é o mesmo: só há uma solução que funciona. Assim sendo, os programas tentam milhões de maneiras diferentes e, de acordo com um conjunto de regras físico-químicas, avaliam as que mais provavelmente correspondem à realidade. As vantagens em usar computadores centram-se essencialmente em dois pontos: primeiro, o facto do preço do material informático ser cada vez mais acessível, e segundo, a velocidade incrível com que se torna possível testar várias hipóteses experimentais.

Infelizmente, apesar de imensos progressos, os computadores hoje disponíveis, assim como o conhecimento actual de física e química biológicas, obrigam-nos a usar certas aproximações que limitam a qualidade final dos modelos. Contudo, há estratégias que permitem melhorar significativamente os resultados destas simulações, ou seja, a qualidade dos modelos. A que mais frutos tem dado, na história recente, é a chamada *modelação integrativa* (tradução literal de *integrative modelling*). Este tipo de modelação utiliza dados experimentais para guiar os cálculos computacionais. É como tentar chegar àquele restaurante novo com ou sem um mapa: com o mapa evita-se a maioria dos caminhos errados, dependendo claro da atenção do condutor. Não é uma estratégia infalível, mas limita bastante a quantidade de modelos errados produzidos pelo computador. Além disso, ajuda também a escolher quais os modelos mais representativos, aqueles que mais próximos devem estar da realidade, uma vez que estes devem concordar mais com os dados experimentais que os restantes.

O que eu tenho feito nos últimos quatro anos, e que se encontra resumido nas folhas

desta tese, é desenvolver novos algoritmos, melhorar alguns já existentes e analisar a estabilidade e performance de métodos desenhados para a *modelação integrativa* de interacções entre proteínas. Na prática, envolve um misto de informática, biologia, bioquímica, física, e matemática. Os seis capítulos da tese descrevem seis trabalhos independentes, a maioria publicada em revistas internacionais e dois dos quais desenvolvidos em colaboração com outras universidades. O **capítulo 1** é uma introdução geral ao estudo da estrutura de proteínas e das suas interacções. O **capítulo 2** aprofunda os conceitos de *docking* e modelação integrativa numa revisão da literatura recente. O **capítulo 3** apresenta um estudo sobre as expectativas a ter quando se produzem modelos de interacções a partir de modelos de proteínas. Os **capítulos 4 e 5**, propõem dois novos métodos, um recuando até Darwin e ao conceito de co-evolução para prever quais os aminoácidos responsáveis pela interacção, e outro recorrendo a técnicas de agrupamento para analisar os resultados das simulações. Por fim, o **capítulo 6** resume os últimos seis meses do meu doutoramento e expõe métodos para simular e modelar interacções que envolvem bastantes proteínas simultaneamente de forma eficaz.

Acknowledgments

Writing a thesis is a one-man effort. Surviving it is not. During these last four years, I was extremely fortunate to be surrounded by some of the smartest and kindest people I have ever met. I could not have made it without each one of you, regardless of how much you think you contributed. To all of you, thank you.

Alexandre. Saying that I owe you whatever career I will have in the future is a gigantic understatement. When I first walked in your office, half an hour later than agreed, I had no clue what I was getting myself into, scientifically speaking. Six years later, I am very happy - another understatement - you had the patience to wait for me. I have really learned some good science during this time, despite never working on my Ph. D project, and I really enjoyed working together with you on all these projects. More importantly, I have to thank you for showing me how to be a good person in a research world that is too often cutthroat and competitive. It is easy to be successful when you are surrounded by smart people; many groups are so. The tricky bit is to make sure the group behaves as indeed, a group, instead of a collection of people. Here at Utrecht, you built nothing short of amazing. A world-leading research group - trying to be humble here - where members are *friends* besides being colleagues. This atmosphere is absolutely correlated with your own character, and I mean a real correlation here, not the ones we usually deal with! In six years, I have never seen you lose your temper, never seen you lose hope, and never heard you criticize somebody unfairly. Instead, you were and are a constant source of support and inspiration to all of us. Even when times are hard, and you had some rough ones lately, with papers being delayed, grants being rejected, and all personal issues added to the mix, you have never let it influence the way you treat people. That is your best quality, on top of course of being cunning, pragmatic, and curious just like a scientist should be. The influence you have on all of us is obvious and it goes without saying that thanks to you, and being humble again, we are a class apart from other doctoral students. I only hope that one day, if I make it to the top of the food chain, I can be half of what you are.

Adrien, Ezgi, and Panos. My two official paranymphs and the unofficial third. What can I say sometimes? In Portuguese we say that 'if you did not exist, someone would have to invent you', but I think in your case nobody could dream of having the audacity and brilliance of making such perfect people. I cannot grasp how much each one of you molded me as a scientist, let alone as a person. Adrien, your patience and kindness know no limits. I started as your student, became your colleague, and ended up your mini-bo... sorry, friend, I mean friend. I cannot imagine, as Panos would say, how I could have survived and thrived as I did without you. I think you do not understand exactly how much you are worth, as a human being, and as a scientist. You are truly one of a kind, and do not, ever, let any funding agency or recruitment process tell you otherwise. Ezgi, my Queen, your dedication and calm are like lighthouses in a storm. Life is tough, we all know it, but to navigate through it with such grace and kindness takes a special personality. Fortunately, I do not need to write everything, you know. I am relieved that despite you moving to another country and another group, we managed to not only keep in touch, but remain *true kankas*. Thank you for all the support during these last months, with writing, with designing, with trying to produce this manuscript while trying to do fifty other things, and perhaps more importantly, with remaining a - sort of - nice person. Concentration, dedication, persistence, and delegation. Did I do it well? Panos, you know why you are the unofficial right? The cycle should have been closed with you, but the rules said explicitly 'paranymphs should not be random'. Man, where can I start? You *are* science. Really, you breathe science, and cigarettes. Thank you for all the inspiration, for all the collaborations, for all the phone calls at random times

since you moved to Germany, and for being such a true friend. I owe so much. You only owe me holidays in Paros. Rolf talked about serendipity. I am very fortunate to have arrived in Utrecht at - roughly - the same time as you three and to have shared all those moments. I doubt I will ever be as fortunate again in life. Thank you, for making me a better person, for making me a (better) scientist. Thank you for everything.

Thanks to these people, and since unlike in chemistry, 'similars' attract, our group is a gold mine of amazing. In no particular order: Mik-mik, I hope I showed you that there is more to Portugal than construction builders. Never change, really. Christophe, you ran away to Australia with Chantel, and now I have a reason to want to go there. Thank you for the voice of reason you very often provided! Marc, as I told you before, you are a box of surprises. There is the Tibetan Zen-ness you practice; and there is the BMW 1200 GS and the power tools you store in the shed you apparently built yourself! I wish you and Cecile all the best, you guys are a fantastic couple. Koen! The youngest, brightest, most promising student… you traitor! Just kidding man, I wish you the best of luck with your Ph. D, you will do a great job for sure. Speaking of youngest, Charleen, princess of the lab. You came to the group as a bachelor student together with Wouter and soon became irreplaceable. You have a big heart and an insatiable curiosity, I hope the ETH matches your expectations! Thanks for the desk sharing over almost one year. Mehdi boy, you are special too, in your own special way. Focus! You need to learn how to focus! Wish you the best of luck in Germany, you can always keep asking me questions via email. Last but not the least, Gydo-dude, Bart the Bard, you are an amazing character. We doubted you at first, coming from FOM and physics and all, but you have proven us all wrong. Thanks for the company, particularly over this last year. The seat of 'senior' is now yours: with great power comes great responsibility! Anna, Li, Cunliang, Stefano, Babis, you are the group now. You have huge shoes to fill but I am sure you will do it, collectively. Good luck! It was a pleasure to welcome you all in the group and I hope it was, is, and keeps being a pleasure to you too.

The computational group was constantly expanding and contracting over the last four years. There are a lot of people I did not mention that were fundamental for my work, my apologies! There are two 'sets' of people I would like to mention specifically: collaborators and students. The first 'set' includes all those with whom I had the pleasure to work with or share my desk with temporarily. Yao and Maarten, the MraY story is a mystery, but I am sure you will crack it! David, you are an honorary member of the NMR group already. Muito obrigado pelas visitas, pela ciência, pela amizade, e força aí no último ano. Beijinho à Liliana também! Fleurtje, thank you for restoring my faith in collaborations. It was wonderful to have you here in Utrecht, sharing not only my office but more importantly, all those sweets and chips. Let's meet on the other side of the ocean soon. Christina, keep dodging those kangaroos! My apologies on behalf of me and David, I hope one day you meet a normal Portuguese person, really. Annin-san, you were supposed to be an Indian student, not a cheerful blonde! You fitted in so well, it was a shame that you left. Good luck over there and arigato gozaimasu. Jeff, I hear you might visit soon! Good! Gosia, keeping it simple, I am sorry I did not have all the time in the world to answer all your questions. I hope I can do it better now that you are back! The second 'set' of people, the students, were one of my greatest joys during my Ph.D. Doing research is fun, but sharing it with others, teaching them what you know and then seeing them thrive is such a fulfillment. In chronological order: Lisette and Richt, it was a complicated project but I could not ask for better 'first students'; Remco & Bas, thank you for so many plots and charts and figures and for the happiness you brought to the group; Sheran and Sanne, Jordi and Bram, you were here for

a super short period, I was mostly in Italy, but I still have your two amazing posters on my wall; Willem & Bertjan, destroyers of papers, guys, I had a blast co-evolving the world and showing it does not… oops!; Max and Cedric, Manisha, not technically my students, but thank you too for the good atmosphere!

Besides the computational people, there are other bunker-dwellers that deserve some praise. First and foremost: Klaartje. What would Adrien do without you? What I wrote about him applies to you, so go back and read it again. I cannot put in words how much I value your friendship and your advice. And Gael, and Elliot, and now Charlie, it is very sad to realize that my French and Dutch are so bad that I have trouble communicating with 3-year olds... They are wonderful kids, a reflection of their parents. Be sure to send pictures of them once in a while! Thank you also for the summary, twice! Murphy...

Second and foremost: Tessa. We fought, we argued, we 'collaborated', and we laughed together. I went nuts often because of you, but I would have gotten nuttier without you. In the end, I am extremely happy we overlapped completely. Excellent student, you deserve all the best in Cambridge. I am sure you just need the right opportunity to become a great scientist. Keep going strong! And keep eating chocolates! And sharing them… with me?

Third and foremost: the Balvincibles. Mark, Markus, Elwin, Mohammed, and others, it was amazing what we did and a shame we waited for so long! Thank you guys for the good environment in the lab, which might or not include wine stains in the ceiling. Markus, you should believe your simulations more and your NMR less, just saying! Erman, I separate you from the bunch for a good reason. It is obvious why Ezgi chose you: your cooking skills. Man, those kofte... You are an incredible kind and smart guy, I really hope you realize your goals and that they bring you closer to Ezgi. Thank you for all the kanka moments, for Turkey, and for my desktop background!

Now on to everybody that I missed. Amanda, thank you for being a good office mate! I know you will become a great teacher, good luck! Eline, thank you for believing I could run 10K. It was indeed a challenge! Good luck EPRing! Hugo, Klemen, and Alma, night-shifters, thank you for the jokes and for putting up with me late at night in the lab! Klemen, you know, relax! Hugo, you know, keep doing your thing, never quit nor change! Alma, too much assignment will kill you! Maryam and Rama, you were here when I arrived and I know it has been a difficult journey, thanks for the company also during the late nights and good luck in the future. Kitty and Sid, bright young 'solid-staters', good luck wherever you go next! Thanks Kitty for the D&D nights, why was it so late? Cecília, you have such big shoes to fill, being the new Portuguese representative. Do not yield to pressure, do not go crazy. ssNMR is an art, more than science (sorry guys!), so do not be afraid to rely on people around you and to pursue your own ideas. Ninguém faz isto sozinho, tem calma, e apoia-te na malta no grupo. Vais conseguir! Hans and Johan. Thanks for putting up with my nitrogen spilling, it was a pleasure to help a bit! Thanks for saving my sub-woofer too Johan! Barbara, thank you for all the support!

Finally to the big shots: Rolf, Gert, and Marc. It takes a special group of people to manage all those special people. Together with Alex, you are the catalytic tetrad that makes the group work. Rolf, you are a source of inspiration. Your knowledge, not only about NMR, is so vast that it makes me doubt the human brain has any storage limit. Thank you for the insights, and also for all the modeling help. The real modeling. Gert, it was a pleasure to join you during the Bioinformatics courses. Among all the chemistry, there has to be a biologist to rule it all and give it meaning. That is you. Thank you for keeping me on the bio side. Marc, the tenacity and furor with which you do science are admirable. Easy? Forget about it, aim

high, aim big. When I arrived, solution was the majority in the group, then it became us, and now the mantle is yours. Good luck!

Interestingly, other people find the NMR bunker attractive and decide to spend some time there or with us. Tania, Martina, Daniel, Hedwig, Deniz, and all the others! The Bijvoet meetings would not be the same otherwise.

Lastly, oh, you thought I forgot you? Thanks for being the worst office-mate ever. I hated every single minute of it, I am so happy I am leaving. Not. Ramoncito, you will make it. You have what it takes, trust me. Thanks for the jokes, they helped so much during the hardest of times.

There is not so much time outside science and the lab. Yet, there were of course moments and people that were always there, in the shadows, always ready to distract me from my problems, share a laugh with, or just go to the movies with. A particular word of praise goes to my two house mates, Bruno and Steven. Bruno, I am really happy I got to know you better than in Coimbra. You were the first person I shared a house with, was it obvious? Foi sempre uma segurança saber que em casa havia alguém com quem mandar umas asneiras para o ar (tão politicamente correcto). Obrigado rapaz, desejo-te o melhor, vai apitando! Stevie Wonder, you were my house mate, but you are also my oldest friend here. You and your family - Rob, Dominique and Vincent - are my *de facto* family here in The Netherlands. Knowing this was the safety pin in many occasions. As always, things change, people change, but feelings remain true. You are like a brother to me. I will always be here for you, wherever here is. I owe you too much. Thank you. Now go back to finish writing your ten grants and fifty papers! To Casper, Ana, Sacha, Marti, Timo, Maja, thank you all for the dinners and beers here and there. It has been 6 years, it was a pleasure! A shout also to Will and Diana, Daniel, Anna, and Ana, for allowing me to have a much-needed non-scientific discussion once in a while. Stay true to yourselves, all of you.

Finally, there is you. You are temperamental, lazy, impulsive, stubborn, and a bit stupid sometimes. On the other hand, you are beautiful, graceful, and I could not imagine my life without you. Of course I am talking about Juko and Yoga, the two cats I share my house with. As for their owner, well, I just cannot imagine this thesis, these four years, and my life at this moment without her. We are very different people but we manage to push each other forward, on to being better people. I have learned so much from you and I have shared so much with you. Thank you Dominika. Or should I say, miau Dominika. Dziękuję bardzo. Kocham cię.

Agora em bom português. Começo pelo Marcelo. Abençoada a hora em que fomos ao Woolloomolloo. Tens sido um apoio gigante, acho que nem vale a pena ir por aí. Obrigado pela sanidade mental nas muitas horas de insanidade durante estes últimos quatro anos. Não te trocava por nada pá! Força aí nos miócitos, mete esses meninos a bater!

Malta do bigode - Paulo, André, João, João, Rui, e Luís - isto é que foi hein? Três casamentos, Dallas, Boston, Coimbra, Utrecht, Barcelona, em breve Bélgica, e Hannover, e centenas de milhares de emails. Vocês são a prova de que a amizade é uma das forças fundamentais do universo, não há distância que a abata. Momentos inesquecíveis, obrigado por tudo! Obrigado pelas visitas também, fomos quase todos a quase todo o lado! Puto JA, 26 anos de amizade, interrompida pelo meio, merece uma menção honrosa. Boa sorte para o próximo passo, espero estar cá ainda a tempo de te visitar, e vice-versa. Um beijinho também às meninas: Raquel, Inês, Marguerita, Sofia, e Su! A manter os machos em ordem desde 2004!

Meninos de Coimbra: João, Eduardo, e Sara. Não vos consigo agradecer o suficiente por tudo. Vocês são parte da razão pela qual estou aqui, pela qual estou são (física e mentalmente),

pela qual estou feliz. A distância custa a suportar, e nem sempre temos tido grande tempo para conversar em condições, especialmente no último ano. Vocês merecem o melhor que há no mundo. Obrigado por serem quem são, e por me terem feito quem sou. Se o sou, devo-o bastante a vocês. Adoro-vos.

Hugo, João. Devo-vos o apêndice, e consequentemente a vida, além de uma enorme dose de lições de vida. Obrigado pela paciência nos últimos anos, pelas visitas, pela amizade. Continuem fiéis a vocês mesmos, é raro encontrar amizade em estado tão bruto e puro.

Por último, família. Pai, nem sei bem como pôr isto em palavras. Os últimos dez anos não foram nada fáceis, nem os últimos quinze, nem vinte, nem quase trinta. Ainda assim, conseguiste manter-me na ordem e guiar pelo meio de tantas contrapartidas. Tenho imenso orgulho em ti. Obrigado. Cidália: as pessoas perguntam-me muitas vezes porque insisto em não te chamar madrasta. Não faz juz ao que representas, não és uma substituta. És o centro da família, és dona de um coração gigante e de uma bondade sem limites. Não sei como teria sido a minha vida sem ti. És a prova de que a familía é construída e não herdada. Pedro, menino, ser irmão à distância é complicadíssimo. Tens cabeça para ser alguém, usa-a, não desperdices as oportunidades que te dão. Acima de tudo, mantém-te fiel aos teus princípios, e continua a ser tão bom puto como és. Avó Milú e Avô Lino, peço-vos imensa desculpa por ter fugido e ter deixado a Skye no vosso quintal. Espero que ela vos tenha trazido um pouco da companhia e da amizade que seriam esperados de mim. Não foi fácil, mas conseguimos os três, e a nossa relação ficou mais forte. Obrigado por nunca deixarem de estar desse lado e de se fazerem ouvir. Leninha, que te hei-de de dizer? Devo-te tanto pela coragem e determinação que me fizeram chegar até aqui. Foste a primeira da família, que eu saiba, a concluir um doutoramento, e contra tudo e contra todos, continuas a lutar sempre por aquilo em que acreditas. És fantástica, e compreendo perfeitamente porque é que a minha mãe te manteve sempre por tão perto. Finalmente, e porque a família não é feita só de laços de sangue, obrigado tia Marília e prima Margarida! Sempre a voz da razão, sempre com bons conselhos. Obrigado também à D. Alice e Sr. Leitão, por serem tão boa influência no meu carácter e na minha educação. Torna-se fácil viver quando se está rodeado de gente assim.

I have more acknowledgments pages than Panos: mission accomplished.

João

> "The **J** is pronounced like the French **J** in *jour*, *bijou*, *jeunesse*. The first syllable is unstressed and therefore the first **o** sounds pretty much like a **w** in English (i.e., a semi-vowel). The stressed syllable is the final **ão** which is a nasal diphthong. You could try pronouncing the **ow** in cow with a nasal sound. It's tricky to get it if you can't hear it - actually it's tricky even if you can hear it. Try saying **Jwaong** quickly remembering what I said about the **J** sound in Portuguese and with the stress on the nasal **a**."

List of Publications

- **Rodrigues JPGLM**, Trellet M, Schmitz C, Kastritis P, Karaca E, et al. (2012) *Clustering biomolecular complexes by residue contacts similarity.* Proteins: Structure, Function, and Bioinformatics 80: 1810–1817.

- Dias DM, **Rodrigues JPGLM**, Domingues NS, Bonvin AMJJ, Castro MMCA (2013) *Unveiling the Interaction of Vanadium Compounds with Human Serum Albumin by Using 1H STD NMR and Computational Docking Studies.* European Journal of Inorganic Chemistry 26: 4619–4627.

- Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, **et al.** (2013) *Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions.* Proteins: Structure, Function, and Bioinformatics 81: 1980–1987.

- **Rodrigues JPGLM**, Melquiond ASJ, Karaca E, Trellet M, van Dijk M, et al. (2013) *Defining the limits of homology modelling in information-driven protein docking.* Proteins: Structure, Function, and Bioinformatics 81: 2119–2128.

- Kastritis PL, **Rodrigues JPGLM**, Bonvin AMJJ (2014) *HADDOCK(2P2I): a biophysical model for predicting the binding affinity of protein-protein interaction inhibitors.* J Chem Inf Model 54: 826–836.

- **Rodrigues JPGLM**, Bonvin AMJJ (2014) *Integrative computational modeling of protein interactions.* FEBS J 281: 1988–2003.

- Kastritis PL, **Rodrigues JPGLM**, Folkers GE, Boelens R, Bonvin AMJJ (2014) *Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface.* J Mol Biol 426: 2632–2652.

- Hopf TA[*], Schärfe CPI[*], **Rodrigues JPGLM**[*], Green AG, Kohlbacher O, et al. (2014) *Sequence co-evolution gives 3D contacts and structures of protein complexes.* Elife 3.

- Ferguson FM, Dias DM, **Rodrigues JPGLM**, Wienk H, Boelens R, et al. (2014) *Binding Hotspots of BAZ2B Bromodomain: Histone Interaction Revealed by Solution NMR Driven Docking.* Biochemistry. In press.

- **Rodrigues JPGLM**, Karaca E, Bonvin AMJJ (2015) *Information-driven structural modelling of protein-protein interactions.* Methods Mol Biol 1215: 399–424. doi:10.1007/978-1-4939-1465-4_18.

Curriculum vitæ

The author of this thesis was raised in Coimbra, Portugal, home to one of the oldest universities in Europe. Born on the 29th of July 1986, he obtained his high school diploma in 2004, from the Liceu José Falcão. In September of the same year, he enrolled in a four-year bachelor study programme in Biochemistry, at the Faculty of Sciences of the University of Coimbra. He completed his studies, in 2008, with a thesis on genomics, and within a month started a Prestige Master in Biomolecular Sciences at Utrecht University, The Netherlands. He graduated in 2010, with a GPA of 4.0, after two research internships, first with Alexandre Bonvin at Utrecht, and then, by the latter's recommendation, with Michael Levitt at Stanford University, USA. Since his graduation, the author has been employed by the NMR department of Utrecht University, conducting research in the group of Alexandre Bonvin towards a doctoral degree.