Sede Amministrativa: Scuola Internazionale Superiore di Studi Avanzati

Settore: Molecular Statistical Biophysics

Scuola di dottorato di ricerca in FISICA
Ciclo XXVII

Assessing the structure of proteins and protein complexes through physical and statistical approaches

Direttore della Scuola : Ch.mo Prof. Guido Martinelli
Supervisore : Ch.mo Prof. Alessandro Laio

Dottorando: Edoardo Sarti

Abstract

Determining the correct state of a protein or a protein complex is of paramount importance for current medical and pharmaceutical research. The stable conformation of such systems depend on two processes called protein folding and protein-protein interaction. In the course of the last 50 years, both processes have been fruitfully studied. Yet, a complete understanding is still not reached, and the accuracy and the efficiency of the approaches for studying these problems is not yet optimal.

This thesis is devoted to devising physical and statistical methods for recognizing the native state of a protein or a protein complex. The studies will be mostly based on BACH, a knowledge-based potential originally designed for the discrimination of native structures in protein folding problems. BACH method will be analyzed and extended: first, a new method to account for protein-solvent interaction will be presented. Then, we will describe an extension of BACH aimed at assessing the quality of protein complexes in protein-protein interaction problems. Finally, we will present a procedure aimed at predicting the structure of a complex based on a hierarchy of approaches ranging from rigid docking up to molecular dynamics in explicit solvent. The reliability of the approaches we propose will be always benchmarked against a selection of other state-of-the-art scoring functions which obtained good results in CASP and CAPRI competitions.

Preface

Chapter 2. Section 2.1 is a review of the methods presented in several works by Miyazawa and Jernigan, [MJ85, MJ96], Sippl [Sip90] and Baker and co-authors [TBM$^+$03, RSMB04, GMW$^+$03]. Section 2.2 contains a summary of the work presented in [CGL$^+$12], here presented with the permission of authors, and expands the concepts hereby mentioned. 2.1 is a Figure from Ref. [CGL$^+$12], copied with the permission of the authors. Section 2.3 is a summary of the methods presented in the works whose citations can be found in the text.
Chapter 3. The work here presented has been published in [SZC$^+$13].
Chapter 4. The work here presented has been published in [SGS$^+$15].
Chapter 5. Paper in preparation.

Table of Contents

Thesis outline

Proteins are a wide class of biological molecules of great importance for the functioning of the cell. They are assembled by the rybosome as unfolded linear chains of amino acids, and in order to perform their functions, they have to assume a well-defined structure that they do not possess at the moment of their assembly. The change from a stretched conformation to a more compact one is called *protein folding*. For small proteins this process is independent of external factors: it is proved that the correct, functional fold in many proteins is only dependent on its amino acid sequence [AHSW61]. This confirms that protein folding is based on mechanisms which are independent of the specific system and can be studied from a physical point of view.

If a protein would have to explore randomly the conformational space until the correct and stable state is found, the folding process would take a time longer than the age of the universe [Lev69]. The fact that folding occurs in much shorter times poses restrictions on the shape of the free energy landscape of proteins, which have to someway drive the system towards the correct (*native*) conformation. A strategy to study the process of protein folding is thus to calculate the free energy of the system. However, an exact calculation of this quantity is far too complex. Thus one has to rely on some kind of approximation and compute some other quantity that mimics the properties of the true free energy.

The same approach can be used to solve a related problem: *protein-protein interaction*. The interaction network among proteins is fundamental for all processes occurring inside the cell: notable examples are metabolic pathways, signaling, channeling, motor skills.

Determining the correct state of a protein or of a protein complex is thus central not only for basic science but most prominently for medical and pharmaceutical research.

This thesis is devoted to devising physical and statistical methods for recognizing the correct conformation in problems of protein folding and protein-protein interaction.

In **Chapter 1** we will present an overview of the structural organization of protein and protein complexes. Then, we will describe the most renowned experimental and computational methods to investigate the protein folding and protein-protein interaction. Particular attention will be reserved to computational methods based on physical laws and statistical concepts. Among them, we will introduce the importance of *scoring functions* and their flexibility in a vast range of different situations.

The subject of scoring functions will then be elaborated in **Chapter 2**. First, we will propose a rigorous definition of scoring function, then we will describe from a historical perspective the development of *statistical scoring functions*, or *knowledge-based potentials* (KBPs), then we will focus on Bayesian Analysis Conformation Hunt (BACH) [CGL$^+$12], a statistical scoring function devised for recognizing the native fold of monomeric proteins. Chapter 3, Chapter 4 and Chapter 5 will be dedicated to the description of improvements of the BACH method and its application to different contexts.

Specifically, in **Chapter 3** we will describe an alternative method to calculate the solvent exposed surface area of a protein in order to account for the interaction between residues and solvent. The new method is based on the Linear Combination of Pairwise Overlaps (LCPO) [WSS99] approach to calculate the solvent accessible surface area (SASA) of a protein. We will describe how a suitably modified implementation of the LCPO methods is able to confer to the BACH algorithm a higher speed, a slight increase in accuracy and the possibility of calculating derivatives of the scoring function with respect to atomic coordinates.

In **Chapter 4** we will describe an extension of BACH scoring function aimed at studying protein-protein interaction (PPI) problems. We will justify the possibility of applying BACH to PPI with the help of a formalism based on information theory. This theoretical framework will also suggest manners to refine BACH parameters in order to increase the accuracy of the score in both protein folding and protein-protein interaction problems. Other improvements will be presented as well: the introduction of a term accounting for steric clashes and the formulation of an accurate procedure to assess the performance of BACH and other state-of-the-art scoring functions on challenging test sets

of protein complexes.

After having assessed the performance of BACH in recognizing the native structure of protein complexes, in **Chapter 5** we will present the results of a procedure aimed at predicting the structure of a complex from the unbound structures of two monomers. Through increasingly accurate refinements, the properties and performances of BACH and other two state-of-the-art scoring functions will be investigated. The results of this study will highlight the connection between the procedure used for generating the structures and the function used to score them, and we will be able to draw some conclusions on the characteristics of both.

Chapter 1

Computational protein structure prediction

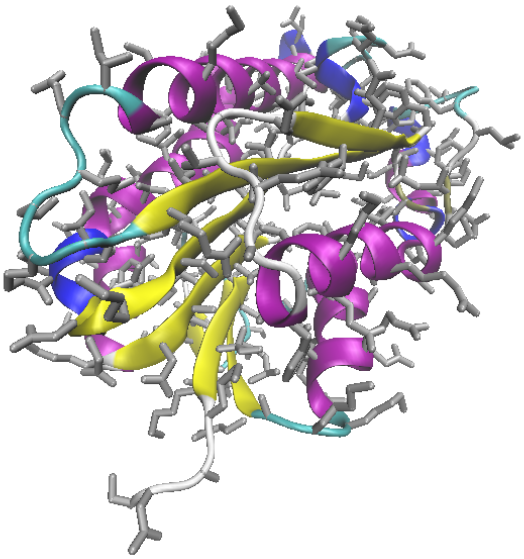

Figure 1.1: A protein rich in secondary structure: α-helices are colored in purple, β-sheets in yellow. The PDB code of the protein is 3D7L.

When observing the geometry of even a small globular protein like the one in Fig. 1.1, one cannot miss to notice how convoluted the network described by its backbone chain is and how apparently intricate and unordered are the interactions between its parts. This complexity stems only from the sequence of the amino acids of the protein, and allows the molecule to perform the function(s) it was selected for. Knowing the structure of an active protein is thus of enormous interest for the biological community and central for pharmacological

studies. Although during the past fifty years experimental techniques made huge improvements in the field of protein structure determination, a consequence of the slowness, difficulty and cost of the process is the very large number of structures that still have to be determined. The ability of predicting the structure of a protein from its sequence of amino acids is thus seen as both a most wanted advancement for all sorts of practical purposes and an important step in understanding the physical process underneath protein folding.
An even more awaited advancement is the ability of predicting the structure of protein-protein complexes. Most of the functions performed by cells are grounded on interaction processes among proteins: well-known examples are signal transduction cascades, transport across membranes, metabolic processes, all of them crucial targets for treating a large quantity of diseases. To predict if two or more proteins will or will not interact with each other through a certain interface is thus a problem taken into great consideration both in the biological and medical fields.

1.1 The geometry of a protein

The intricate geometry of proteins shows three different levels of organization, called primary, secondary and tertiary structure.
Under the name of *primary structure* goes the information about the distinctive sequence of amino acids a protein is composed of. Knowing the primary structure of a protein means knowing its topology, the information that a chemical formula of the compound provides. Rigorously speaking, amino acids can undergo many chemical reactions: in presence of catalyzers, ions, acidic or basic solvents, some of their side chains can lose or gain protons. Proton transfers often cause the start of cascade reactions which most of the times are of great importance in order to carry on the main functionalities of the protein. It is also common to observe hydrolization, phosphorylation or sulfur bridge breaking processes, and this only to state some of the most important categories of chemical reactions involving sidechains. Chemical reactions happen also in the solvent surrounding the protein: water molecules at thermodynamic equilibrium and at room temperature undergo autoionization with a certain probability. Although these events may be very important from a chemical perspective, they are rarely relevant in the framework of structural prediction. Thus, we will ignore them. For our purposes, we will assume that the chemical

formula of a protein, and thus its primary structure, remains constant.

When assembled, a protein is a completely stretched and untangled chain of amino acids. Although a non-negligible fraction of proteins partially maintains the initial lack of shape, most of them assume less flexible and more ordered conformations. Locally, they can fold in a helical fashion (the majority of which are *α-helices*) or can assume the shape of a slightly twisted sheet (*β-sheets*). What these structures have in common is the spatial locality. They are identified by the name of *secondary structure*. In the functional (namely, *native*) conformation of proteins its presence can vary considerably, ranging from 10% to as much as 90%. Secondary structure is the protein's intermediate level of structural organization, and is stabilized by hydrogen bonds created by the backbone of the protein, which is the sequence of amide groups from which the side chains (namely, *residues*) stem.

The secondary structure is organized in domains which can be connected by what is called *tertiary structure*. This level of organization reflects the global network of interactions which provide to the protein its final and stable form. Even if the matter is still somehow debated, it is accepted by the majority of the community that the main driving force for the creation of both the secondary and tertiary structure is the degree of polarity of the side chains. 8 over the 20 types of residues are indeed hydrophobic, and 4 have hydrophobic regions due to aliphatic carbons. Since water, which is by far the most common solvent molecule, is polar, the hydrophobic parts are subject to a force which pulls them together in the bulk of the protein, minimizing their exposition at the protein-water interface. This implies that the stability of both secondary and tertiary structure can be modified even without changing the chemical composition of the protein. For example, most of the proteins whose native conformation is ordered and stable lose much of their structural complexity when immersed in a solvent with both hydrophobic and hydrophilic parts, like ethanol or methanol. The hydrophobic part of the solvent will weaken the interaction that stabilizes the tertiary structure, while the hydrophilic part will compete with the solvent-solvent hydrogen bonds.

1.2 Systems at equilibrium

The nature of the solvent is indeed not the only factor that can alter the stability of a protein. The protein-solvent system is prone to change its properties in relation to changes in macroscopical observables like density, temperature or pressure. Usually, an environment at constant temperature, pressure and number of molecules is considered. This arrangement is defined as *closed system*, meaning that the system cannot exchange matter with the environment, but can exchange energy to keep some macroscopical quantities fixed. If these quantities are temperature and pressure (or, sometimes, volume), the system is at equilibrium and the set of different configurations the system can take is called *canonical ensemble*.

In conditions of equilibrium, several proteins unfold and re-fold spontaneously due to thermal fluctuations [HGMO$^+$06]. The conformations contained in the canonical ensemble still differ by secondary and tertiary structure arrangements. Such sets of conformations are the object of interest in Biophysics. Statistical mechanics offers powerful tools for the investigation of the properties of the system: Gibbs's and Helmoltz's free energies have the property to be minimal for the equilibrium conformation. Moreover, systems at equilibrium are easier to study experimentally, hence estimates of their properties are more accurate. Finally, the canonical ensemble of a protein at a specific temperature and pressure can be used to recover all the physical quantities that characterize the system, such as its free energy landscape, its entropy and the probability of the system to be trapped in a metastable state.

1.3 Protein complexes

There is yet another level of complexity which involves proteins: they can form complexes, and organize in what is called *quaternary structure*. Interaction between proteins occur when two or more monomers bind together to carry on their biological function. Proteins have been observed to interact in groups of very different sizes. As an example, many transduction paths are operated by complexes of only two subunits (dimers), while very large molecular machines are usually found performing motor tasks like DNA replication [BB11], cargo [Sch04] or motion [KB04]. Protein folding and protein-protein interaction do not necessarily happen separately: there are families of proteins whose unbound

state is partially unstructured and which complete their folding process upon binding. In this case, we talk about *obligate* complexes, and usually they are characterized by a much longer lifetime [NT03]. There is still much controversy on the mechanism of binding for obligate complexes. Our attention will be mainly focused on complexes formed by monomers which are already folded at the moment of the interaction (*non-obligate* complexes).

1.4 Structure determination by X-ray and NMR spectroscopy

Given the high degree of complexity of the three-dimensional structure of proteins, the process of recovering it from experiments is also very lengthy and complex. Among the many techniques used for this purpose, two of them proved particularly successful during the years: X-ray diffraction [KBD$^+$58] and nuclear magnetic resonance (NMR) spectroscopies [Wut01]. The advantages and drawbacks of these two approaches are usually believed to compensate.
X-ray spectroscopy provides an accurate estimation of the positions of all the atoms except the hydrogens, but the protein needs to be condensed into a crystal. The process of crystallization can modify the structure of the protein, that will not be exactly in its native conformation anymore. It is usually taken for granted that at least for completely folded, globular proteins, the crystallization process does not alter much the shape of the protein, and that for this class of proteins the crystallized state is thus close to the true native state. Nonetheless, the crystallization process can be extremely time consuming and entails the most relevant limitation of the technique.
If X-ray spectroscopy provides accurate tridimensional models at the cost of crystallizing the protein, in *NMR spectroscopy* it is sufficient to isolate the desired protein in a solvent. The drawback is that this technique allows reconstructing the structure only by algorithms which use the information enclosed in the spectra to recover distances between atoms. Provided there are enough constraints, the problem of guessing a tridimensional structure from distances between atoms has an exact solution. If the number of sufficient constraints is not exactly reached but only approached, the conformational space usually reduces enough to allow for reasonable guesses. Still, the algorithms used in this procedure are complex, and work only for proteins of moderate size.

1.5 Computational approaches to structure prediction

Both X-ray and NMR spectroscopy require costly machineries. The experimental procedures are complex, long to perform and usually offer no guarantees of success. Smaller costs in time and resources is what makes statistics-driven computational approaches competitive with these well-established techniques. Computational techniques do not involve the direct observation of a biomolecule, rather they try to guess the structure of a protein based on the information contained in the primary sequence. Indeed, the seminal studies performed by Anfinsen in the 1960s [AHSW61] prove that the secondary and tertiary structure of a protein only depend on its amino acid sequence. This implies that the knowledge of the primary sequence is in principle sufficient to recover the native tridimensional structure. In this section we will provide an overview of the most important computational methodologies to approach the problem of structure prediction.

1.5.1 Structure prediction by Bioinformatics

A remarkable progress in structure prediction was made by comparative modeling methods, which take as reference experimentally resolved structures. The two most successful methodologies are homology modeling [FS03, MDT09] and protein threading [XLX04, Zha08].

Homology modeling is based on the assumption that proteins whose sequence show similarities are prone to show structural similarities as well. Thus, from a database of resolved structures, a library of matches between sequences and structures is created. To make a prediction on a target protein whose structure is unknown, its sequence is aligned with the ones in the database. If a match in sequence is found, the structure of the protein is modeled on the equivalent protein in the database. Three groups of methodologies can be identified: rigid body assembly [Gre81], segment matching [Lev92], and modeling by satisfaction of spatial restraints [SB93].

In *rigid body modeling* algorithms the structure is assembled from a small number of rigid bodies obtained from the core of the conserved regions [Gre81]. The non-conservative parts are then rebuilt.

The *segment-matching* methods divide the sequence into small parts and then model each of them against a database of known fragment structures [Lev92].

Unlike rigid body modeling, this method requires only a minimal degree of similarity between the known structures and the unknown one [SB93].
The *spatial restraints* methods derives a set of restraints from the sequence alignment and then builds the model which minimizes the number of violations of these restraints.
While homology modeling aligns only sequences, and thus needs to have in the database a sequence homologous to the target one, *protein threading* also works with three dimensional structures [Zha08]. From a database of resolved structures, a library of different protein folds is derived. The target structure is then fitted to each fold in the library, allowing also for insertions and deletions. The quality of each possible fit is calculated by a scoring function: the fold with the highest quality will be the actual structure prediction for the target sequence.

1.5.2 Scoring functions

The description of the techniques mentioned above allows stressing from the very beginning the importance of scoring functions in the framework of protein structure prediction. Algorithms that perform a selection over a large set of possible models need some sort of decision-making step. *Scoring functions* are functions used to select over a set of possible outputs the ones with the best features. They can be divided into two categories: physical potentials [SR95, HSL95, YD96] and knowledge-based potentials [MJ96, Sip90, SKHB97].
Physical potentials try to reproduce the free energy of a conformation by considering actual physical terms combined through a limited quantity of optimized parameters. An important example are the force fields for atomistic simulations, that will be described in Section 1.5.6. The advantage of this kind of approach is that they can usually provide clearer insights about the systems studied. However, due to their explicit formulation, the complexity of the interaction they can describe is limited by the functional form of the terms used to model them.
Knowledge-based potentials instead possess a larger quantity of parameters which are made to capture the relevant physical and chemical details through a learning procedure. Since they do not contain explicit terms to model the interactions, they can account for much more complex effects, whose limit is only the statistical method used to estimate the value of the parameters. This comes at the cost of losing a clear insight on what the potential is actually

describing.
The category of knowledge-based potentials (KBPs) will draw our interest during this work. Tanaka and Scheraga, followed by Miyazawa and Jernigan, introduced the method as a quasi-chemical approximation [TS76, MJ85], while Sippl as a of potential of mean force [Sip90]. There have been controversies on the use of the denomination "potentials" for these methods, as they do not really provide meaningful free-energy estimates [TD96, BN76]. Thus, a redefinition in terms of Bayesian probabilities was proposed by Baker et al. in the Rosetta method [SKHB97], and is now commonly accepted. We will present the concept in more detail in Chapter 2, where we will apply them to other kinds of problems. Further developments in the construction of scoring functions were devised in more recent times, many of them including system-specific terms [DSZ$^+$12]. Notably, the introduction of coevolution is proving very effective in determining patterns of conserved native contacts [CT15, LT13, MPL$^+$11]

1.5.3 Structure prediction of protein complexes

Interaction between proteins is such an important process that during the last decades a large fraction of the computational community's attention has shifted from single proteins to protein superstructures. Indeed, protein-protein interactions are the target of the vast majority of drugs, and are thus of great importance for pharmaceutical studies. Limiting to the case of a dimer composed of two already structured monomers, the two most important issues are the following: understanding if two proteins interact, and, if they do, which is the quaternary conformation adopted by the complex.
The first problem was studied from different perspectives, such as biochemistry [Gol02, HCK02], quantum chemistry [QDWL11], signal transduction and metabolic or genetic networks [SUS07]. Tools offered by bioinformatics were used as well: again, a variety of sequence homology methods were developed. Among them, genomic-based approaches [SUS07] integrate the data from different methods to build an interaction network. Then, they try to predict the function of the complex from the analysis of the network structure. These methods are the ones which brought to the discovery of hundreds of multiprotein complexes, but unlike homology-based models they do not provide hints on the structure of the complexes.

1.5.4 Docking algorithms

Docking algorithms [KKSE+92] try to predict the native conformation of a protein complex by generating a large set of different poses which are then evaluated by a scoring function. The first algorithms for macromolecular docking were devised in the late '70s and were limited to assess the shape complementarity of the possible interaction sites [JW78]. In the early '90s more structures were determined and the available computational power increased substantially: these premises led to the birth of the first algorithms able to perform large-scale conformational searches. However, even nowadays sampling the whole conformational space of two interacting subunits remains unfeasible. A thorough scan of the whole space is possible only if a *rigid docking* technique is applied - that is, only if we consider the two subunits as rigid bodies and we generate the poses only applying rotations and translations. These algorithms are usually accelerated by the use of Fast Fourier Transforms (FFT) [KKSE+92, MRP+01], which enable the program to evaluate at once all the conformations separated only by a translation, or by Geometric Hashing [WN00], which also allows to bring the computation cost of the conformational search from $O(n^4)$ to $O(n^3)$, where n is the number of putative sites. Other approaches include the use of Boolean operations [PKWM00] and genetic algorithms [GWA01]. Many of the state-of-the-art docking algorithms use rigid docking in a first moment, and then select a subset of conformations whose structure is refined. The selection is usually operated by scoring functions which are based on additional information such as predictions of functional sites [ZS01], estimations of the free energy of binding [NL01] and additional structural data [Clo00]. It is debated if rigid docking offers a reasonable starting point to model protein-protein interactions among every couple of ordered, globular monomers. In the case of intrinsically disordered subunits, or in case of folding upon binding, the method clearly lacks the necessary precision.

It is rather well assessed that the regions close to a binding site have often a flexible conformation [LF00], and that the binding site is likely to undergo conformational changes upon binding in a mechanism denominated *induced fit* and postulated in the 1950s [GLC94, Kos58]. *Flexible docking* largely increases the dimension of the conformational space by allowing some of the torsional angles of the monomers to move. After a coarse conformational search, flexibility is allowed on the most promising configurations through a variety of meth-

ods such as ensemble analysis [BRW95], normal mode analysis [ESDW+08], molecular dynamics [DNRB94] or essential dynamics [MR05]. In most cases flexible docking methods provide better conformations, but they have to rely on internal scoring functions to focus on the most plausible interaction sites.
In short, some progress has been made through the years in building docking algorithms, but much more work is needed to provide a quantitative picture of the binding interaction sites and affinities of two monomers, as we will see in Chapter 5.

1.5.5 Competitions for prediction and scoring of native structures

CASP (Critical Assessment of protein Structure Prediction) [CKT09] and CAPRI (Critical Assessment of PRedicted Interactions) [JHM+03] are two community-wide experiments to judge the performances of methods for structure prediction and structure quality assessment. While the former focuses on problems of protein folding, the latter is dedicated to protein-protein (or protein-nucleic acid) interaction problems. The two competitions follow the same procedure: first, an unpublished native structure is communicated to the organizers. In CASP, the amino acid sequence of the structure is supplied to the participants, who will then try to predict the three-dimensional structure of the protein. In CAPRI, either the bound/unbound conformations of the monomers is supplied, or the amino acid sequence and some homology data. Each predictor can send a small number of guesses to the organizers. The guesses are then evaluated with the help of structural estimators: RMSD, interface-RMSD, ligand-RMSD, fraction of native contacts. As a second part of the competitions, the predictions submitted by the participants are shuffled and given to another group of participants which will try to score the nearest-native structure in the set. The difficulty of the scoring competition also depends on the quality of the predictions.
The sets of predictions from CASP and CAPRI experiments constitute a very challenging trial for methods of structure quality assessment, and are widely used for test purposes throughout the community. We will make use of these resources in Chapters 2, 3 and 4.

1.5.6 Molecular Dynamics

Up to now we presented methods to infer the native structure of a protein or of a protein complex based on probabilistic considerations. Although statistical methods are a powerful and efficient tool to treat a large class of problems, they suffer from several limitations. First, they need to rely on a solid statistics, which may not be available, or may be incomplete or biased. As an example, in the Protein Data Bank many classes of proteins are highly underrepresented: although transmembrane proteins are thought to constitute roughly 40% of the different types of proteins present in a cell, only 2% of the PDB is composed by them. Second, we will see in the rest of this thesis that estimators to judge the quality of a model are seldom exact. Lastly, probabilistic methods based on information obtained from structural datasets are not able to investigate on all sorts of quantities and mechanisms connected with the dynamics of the system.

One manner to address these issues is to rely on a physical rather than on a statistical approach, and simulate the dynamics of the system by applying Newton's equations of motion. The method is called molecular dynamics (MD), and has been for decades a central instrument of inquiry in biophysics as well as in other branches of science. We briefly present the advantages and disadvantages of relying on this approach by relating to our focus: recovering the correct structure of a protein or of a protein complex. Indeed, one method to tackle the problem is going through the core biological process, by simulating the folding of the protein starting from a stretched conformation. MD operates by applying the equations of motion and makes the system evolve in time until the equilibrium conformation is reached. If the system was evolved with respect to the correct Hamiltonian, this technique would produce exact results. QM approaches, where both the nuclei and the electrons are treated with the correct Hamiltonian, can be afforded only for small molecules. Among the many approximations that can be considered, the Born-Oppenheimer one is often used, assuming that the relaxation time for the electrons is much faster than the one for the atomic nuclei. Thus the Schrödinger equation for the electron system is solved at any time by considering the external field generated by the nuclei as frozen. The energy of the system can thus be inferred by only considering the movement of the electrons, and will be dependent on the position of the atomic nuclei. Despite the greater simplicity of this method, it still results in very expensive calculations, limiting its usefulness only to

study systems of tens of atoms for as little as tens of picoseconds at most. Calculations made through the density functional theorem (DFT) can be set up for specific systems containing hundreds of atoms, which nowadays are the limit for QM calculations. This is not sufficient for simulating a protein immersed in a solvent, which can contain hundreds of thousands of atoms. This is the reason why classical rather than quantum Hamiltonians is usually exploited. The elimination of the electronic degrees of freedom enables enormous savings of computational time, allowing the simulation of fully hydrated biomolecules. Atoms are generally treated as hard spheres and chemical bonds are approximated by steep parabolic potentials. Clearly within this description, the system cannot modify the chemical bonds during the run. The chemical formula of the compound is thus bound to remain identical for the whole length of the simulation. This restraint might seem exceedingly strict, and indeed it is if one wants to study processes involving chemical reactions. On the other hand, if the subject of the study is protein folding or protein-protein interaction, this constraint helps limiting the already very large conformational space of the system.

Passing from a QM to a classical description, one has to choose an empirical form of the Hamiltonian, which is normally approximated with a sum of analytical terms describing the chemical bonds, Van der Waals and electrostatic interactions. The set of terms and parameters used to describe the energy of the system is globally called *force field* (FF). The FF parameters are often fitted on the potential energy evaluated with QM approaches in smaller systems representing typical parts of the greater systems. In fifty years of development, molecular dynamics passed from simulating thousands of atoms for tens of picoseconds to simulating hundreds of thousands of atoms for milliseconds, with the use of the most powerful state-of-the-art supercomputers [SGB+14]. By using molecular dynamics, the folding mechanisms of many specific systems have been understood, and the structure of some small proteins have been predicted [LLPDS11]. Furthermore, many other phenomena like transitions of ions through channels or enzymatic reactions have been studied [KTS+14]. Disregarding all quantum effects is though proved to be at times a too serious assumption [LLMP+12]. Water, for example, has a prominent quantum nature which results in its extraordinary properties, which are of utmost importance in biochemistry and for all life-related events in general. Although great efforts

have always been made to reproduce the important characteristics of water (i.e. viscosity, surface tension, formation of clathrate structures, condensation and evaporation temperatures, etc.), the parameters used in classical FFs to describe a water molecule are simply too few to reproduce all these properties. The compromise is usually choosing some of the feature the solvent is required to have and treat the others less carefully. Many of the best known force fields are tuned in order to reproduce accurately the dynamics of proteins in a folded configuration. Indeed, folded structured proteins represent the class that can be best observed by experimental techniques, and hence the systems on which we have the most stringent measurements.

Molecular dynamics provides a very rich information. The position and velocity of each atom can be tracked and analyzed. This incredible level of accuracy is compensated by a manageable, but still high time cost: nowadays, a common workstation is able to simulate a few nanoseconds per day of dynamics of a typical system of tens of thousands of atoms. Unfortunately most of the interesting conformational changes usually take place in much greater lapses of times, ranging from microseconds to minutes. Methods for distributed computing were devised [SP00] to exploit the unused cpu times of personal computers. The approach turned out to be very successful. However, much of the computational power is still employed to simulate uninteresting movements of the system, such as the fluctuations of the solvent. In order to cope with larger systems or with larger time scales, there are two approaches which can be taken. The first is making the interesting events happen sooner, or more frequently. The methods going in this directions are called *enhanced sampling* techniques [KRB$^+$04, KRB$^+$02, SO99, LP02]. The second is sacrificing the atomistic detail of the description, and grouping different atoms parts which will move rigidly. This strategy is called *coarse-graining* [MRY$^+$07, CPD12].

1.6 Statistical mechanics and probability

During this quick introduction to the problem of protein structure prediction we made frequent use of concepts and methods derived from two distinct branches of physics. Statistical mechanics is the chief theory beneath molecular dynamics simulation techniques, as well as an important reference for many experimental procedures. On the other hand, the approaches aimed at scoring large sets

of structures, or at predicting the structure based on sequence similarity, are mostly rooted in probability theory. The two fields deepen their roots in seminal works appeared two-three centuries ago, and ran parallel ever since. Although statistical mechanics, especially in its application to quantum theories, makes constant reference to the concept of probability distributions, the methods and the justifications of the theory seldom base on the concepts of probability theory [Sco00]. During the past century, there have been notable efforts to connect the two disciplines: for example, maximum entropy methods have been applied to get rid of the necessity of the ergodic theory as a foundation for statistical mechanics [Cat08]. Nonetheless, the results of a combined vision do not entirely convince the community, mostly because they bring forth hard philosophical issues, such as the definition of physical events as subjective (*Bayesian*)[Jay03] instead of objective (*Humean*) [How00]. As it happens in other branches of physics, there is thus a problem of reluctancy in merging two theories which though seem to converge more and more. The arguments treated during this work will often need to refer to both environments. Not to take for granted a unification which has not been yet achieved, we will limit ourselves to consider the two theories as separate, and to work by analogy when passing from one to another. The author hopes that the general spreading of machine learning and statistical methods in the context of biophysics will bring the community towards a comprehensive view of these ever closer areas of knowledge.

Chapter 2

Methodological background

This thesis is devoted to the improvement and quality assessment of BACH, a scoring function for protein folding and protein-protein interaction. We will thus start by introducing formally the concept of scoring function. Then, we will introduce the original formulation of BACH and describe the procedure to assess its performance in protein folding problems. Since one of the main aims of this work is to extend BACH to the treatment of protein-protein interaction problems, we will then describe some of the state-of-the-art scoring functions which proved to perform well in this task.

2.1 Scoring functions

A scoring function can be defined in a very general way. Consider a finite set of objects $E = \{c_1, c_2, ..., c_n\}$ and a finite set of parameters $P = \{p_1, p_2, ..., p_m\}$. A *scoring function* $s : E \to S \subset \mathbb{R}$ is a function from E to a finite subset of real numbers $S = \{s_1, s_2, ..., s_q\} \subset \mathbb{R}, |S| \leq |E|$, such that the "best" object $c_B \in E$ maps into a predefined extremum of S, $s(c_B; p_1, p_2, ..., p_m) = \text{extr}(S)$.

For this definition to be flawless, there must always be the possibility of defining the concept of "best" object, in order to always have such object in any possible set. As an example, we can consider the problem of rating the quality of a business company. There is no unique definition of what the "best" business company should be. The evaluation criteria may change substantially: one could be interested in the ability to repay credit, or the ability to improve sales, or again the ability to expand its market, or a combination of these and many other factors. Independently of the criteria for choosing what "good"

means in a certain context, there should always be the possibility of finding the best object in any set. Following the example, we cannot define the best company as the one which gives the highest amount of shares, provided that these are at least $ 1.000.000 worth, because there may not be any company in the set satisfying the condition. More specifically, if there are at least two companies which do not fulfill the requirement of distributing $ 1.000.000 worth shares, there will be at least one set in which one cannot tell which is the "best" company: the one containing only those two companies.
If a "best" object can always be defined, it follows that for each set of objects there is only one possible rank. This rank is iteratively built by identifying the "best" object of the set, then taking the subset containing all the objects in the original set except the one selected as "best" and identifying the "best" among them, and so on. If a "best" object cannot be defined for each possible set of objects, the definition still holds, but the map between a certain set of objects and their rank is not bijective anymore: there could be more than one possible rank, or there could also be none. We will see this is the case for free energy scoring functions.

As we saw in the previous example, the definition of what is "best" depends on what one wants to use the scoring function for: this is the reason why scoring functions can be applied in so many different situations. One could also rewrite the definition of a scoring function in the formalism of statistical mechanics: let E be a finite set of configurations c_i of a system taken from the canonical ensemble at fixed $T = T_0$ and $P = P_0$ (or alternatively $T = T_0$ and $V = V_0$). Let there be a finite set of parameters $P = \{k, \epsilon_0, e, h, ...\}$. If we define the "best" conformation as the one dominating the equilibrium ensemble, the definition of scoring function becomes the definition of the Gibbs' (or Helmholtz's) free energy. It is not granted that the equilibrium conformation be present in every possible finite subset of E. Thus, this definition does not imply that for each subset a unique rank exists. In order to define a rank in all the conditions, one has to define the "best" conformation as the "closest" one to the equilibrium conformation. Nonetheless, this introduces the problem of defining a proper distance between two conformations of the same system. In the case of protein folding and protein-protein interaction, one can simply choose the root mean square distance (RMSD), or some analogous function. We will principally use the RMSD for protein folding and the iRMSD (interface-RMSD) for protein-

protein interaction.

Up to now we have introduced the concept of scoring function in a very general and abstract manner, in order to stress that protein structure prediction is just one of the many fields to which one can apply the same concept. In the context of specific interest for this thesis, scoring functions can be divided into two categories:

- *physical scoring functions*, which employ actual physical terms in order to estimate the score of a conformation. In these methods, the parameters are usually fitted over a learning database, but still conserve an appropriate physical meaning. Well-known examples are the force fields used in molecular dynamics.

- *statistical scoring functions*, or *knowledge-based potentials* (KBPs), which are not based directly on physical premises. The physical and chemical information is stored in a large set of parameters in the course of a learning procedure. The parameters do not have a simple physical meaning, and are only values in which the information is stored. A notable example is the Miyazawa-Jernigan potential [MJ85].

The two categories are to be conceived as fluid: indeed, there are scoring functions which combine features from both of them. For example, in the next section we will describe the Rosetta potential [SKHB97], which is a KBP relying on many different statistics, each of which accounting for a different type of physical interaction. We will now introduce KBPs in their historical perspective, which will help us understanding the roots of the theoretical formulation of the BACH algorithm.

2.1.1 Potentials of mean force?

Tanaka and Scheraga [TS76] were the first ones to extract effective interactions from frequencies of contacts in X-ray resolved structures. Miyazawa and Jernigan then proposed a first formalization of the theory estimating contact interaction potentials by means of the *quasi-chemical approximation* [MJ85]. The approximation is based on the assumption that, for residue types a and b,

$$\frac{\bar{n}_{ab}^2}{\bar{n}_{aa}\bar{n}_{bb}} = e^{-e_{ab}} \tag{2.1}$$

where $\bar{n}_{ab}$ is the average number of contacts between residue types a and b, and e_{ab} is the contact energy associated to that couple of residue types, and is a parameter of the model. In their work, Miyazawa and Jernigan also include the interaction with the solvent, which is treated as an extra residue.
Further developments of these concepts brought Sippl to justify KBPs based on probability distributions of pairwise distances [Sip90]. His main contribution was defining the concept of reference state, which will influence the development of KBPs for decades. The reference state is a state of a hypothetical system with which one compares the actual system. Usually, the reference system does not include interactions, so that by calculating the free energy difference between the state of the actual system and the state of the reference system, an estimate of the interaction energy is provided. This is typically done through the inverse Boltzmann formula:

$$\Delta F_{ab}(r) = -kT \left(\log \frac{P_{ab}(r)}{Q_{ab}(r)} + \log \frac{Z_P}{Z_Q} \right) \tag{2.2}$$

where a and b are types of atoms or residues and r is the distance between two units. The $\Delta F_{ab}(r)$ are the parameters stored in the KBP which then, depending on the contact occurrences in the test structure, have to be summed in order to compute the score (a free energy estimate). During the learning phase, $P_{ab}(r)$ is estimated from a database of known structures, and $Q_{ab}(r)$ is estimated with respect to a reference system. One of the first reference systems treated the residues as if they were in gas phase inside a box. Many times, the probability density Q does not even depend on the contact distance r.
This approach is justified saying that the *potential of mean force* (PMF) $W_{ab}(r) = \Delta F_{ab}(r)$ is an estimate of the reversible work required to bring two particles (or residues) of type a and b from infinite distance to a distance r. However, most of the times this assumption does not apply [BN76]. For this reason, the approach by Sippl was long debated and the definition of PMFs was found not appropriate.

2.1.2 Rosetta and the Bayesian approach

An approach based on Probability Theory was proposed in a seminal work by Baker and co-workers [SKHB97]. We can consider the score associated to a specific conformation as the probability $P_{\text{test}}(\mathfrak{C}|\mathfrak{L})$ of finding a certain structure

$\mathfrak{C}$ given a sequence of amino acids $\mathfrak{L}$. Applying Bayes' theorem we get

$$P_{\text{test}}(\mathfrak{C}|\mathfrak{L}) = \frac{P_{\text{D}}(\mathfrak{L}|\mathfrak{C})}{P_{\text{D}}(\mathfrak{L})} P_{\text{D}}(\mathfrak{C}) \propto P_{\text{D}}(\mathfrak{L}|\mathfrak{C}) P_{\text{D}}(\mathfrak{C}) \tag{2.3}$$

where the subscript "test" indicates the probability distributions of the unknown structure to score, while "D" the ones calculated on a database of known structures. As always when using the Bayesian probability, caution must be taken in considering labels. While the final probability distribution $P_{\text{test}}(\mathfrak{C}|\mathfrak{L})$ refers to the structure which is being scored, the likelihood is estimated on a database of known structures. But the Bayes' theorem alone is nothing but the identity $P(A|B)P(B) = P(B|A)P(A)$, where all four probability densities refer to the same system. Thus, by Bayes' theorem alone, we should have written $P_{\text{test}}(\mathfrak{C}|\mathfrak{L}) = P_{\text{test}}(\mathfrak{C}) P_{\text{test}}(\mathfrak{L}|\mathfrak{C}) / P_{\text{test}}(\mathfrak{L})$. This implies that the first and most important assumption of the method is that the likelihood calculated on the test system can be approximated by the likelihood estimated on the database of known cases:

$$P_{\text{test}}(\mathfrak{L}|\mathfrak{C}) \approx P_{\text{D}}(\mathfrak{L}|\mathfrak{C}) \tag{2.4}$$

This is a typical assumption in Bayesian approaches: indeed, the choice are usually either to approximate the likelihood with an analytical formula (a tentative probability distribution satisfying the properties of interest) or to compute the likelihood on another system. If we further assume that the likelihood can be approximated by a product of pairwise probabilities, we have

$$P_{\text{D}}(\mathfrak{L}|\mathfrak{C}) \approx \prod_{i<j} P_{\text{D}}(a_i a_j | r_{ij}) \propto \prod_{i<j} \frac{P_{\text{D}}(r_{ij}|a_i a_j)}{P_{\text{D}}(r_{ij})} \tag{2.5}$$

where i and j are indexes over the atoms or residues considered and a_i and a_j are the type of atoms (or residues) i and j. Eq. 2.5 eventually recovers the Sippl's formula of PMFs. Rosetta method exploits two different databases: while the database used to calculate the likelihood $P_{\text{D}}(\mathfrak{L}|\mathfrak{C})$ is a collection of crystallographic structures, the prior $P_{\text{D}}(\mathfrak{C})$ is calculated on a database of fragments of proteins, with the following assumption:

$$P_{\text{D}}(\mathfrak{C}) \approx \prod_{i<j} P_{\text{D}}(r_{ij}, \theta_{ij}, \phi_{ij}, \omega_{ij} | ss_i, ss_j) \tag{2.6}$$

where $r_{ij}, \theta_{ij}, \phi_{ij}, \omega_{ij}$ are the distance and the angles describing the relative orientation of local structure elements ss_i and ss_j. By writing the learning and

scoring, one obtains

$$
\begin{aligned}
P_{\text{test}}(\mathfrak{C}|\mathfrak{L}) &= \frac{P_{\text{D}}(\mathfrak{L}|\mathfrak{C})}{P_{\text{D}}(\mathfrak{L})} P_{\text{D}}(\mathfrak{C}) \propto P_{\text{D}}(\mathfrak{L}|\mathfrak{C}) P_{\text{D}}(\mathfrak{C}) \\
&\approx P_{\text{D}}(\mathfrak{C}) \prod P_{\text{D}}(a_k a_l | r_{kl}) \\
&\propto \frac{1}{2} \prod_{ss_i \neq ss_j} P_{\text{D}}(r_{ij}, \theta_{ij}, \phi_{ij}, \omega_{ij} | ss_i, ss_j) \prod_{k<l} \frac{P_{\text{D}}(r_{kl}|a_k a_l)}{P_{\text{D}}(r_{kl})}
\end{aligned}
\tag{2.7}
$$

During the years Rosetta scoring function was further improved by the implementation of additional terms to account for different types of interaction. However, the improvements regarded mainly the implementation of the likelihood and the prior probability, while the formalism presented in this section was kept unchanged, and provides a sound explanation of Sippl's PMFs from a probabilistic point of view.

2.2 BACH: Bayesian Analysis Conformation Hunt

Bayesian Analysis Conformation Hunt (BACH) is a statistical scoring function for protein folding based on residue-residue contacts and residue-wise exposure to the solvent. The scoring function is described in its original form in an article by Cossio and co-workers [CGL+12], and more extensively in Cossio's Ph.D. thesis [Cos11]. Another description can be found in Zamuner's Ph.D. thesis [Zam15], where again the method is presented in its original form, and where some modifications are proposed. We will here summarize the statistical method used in BACH scoring function in the light of the formalism described in Section 2.1. Then we will expose the main results of the original BACH method, along with the tools used to assess its performance with respect to other state-of-the-art scoring functions.

2.2.1 BACH statistical method

BACH is a residue-wise knowledge-based potential founded on a Bayesian formalism inspired to the one described in Section 2.1.2. The ideas at the basis of the construction of BACH scoring functions can be summarized in the following:

- The score of a structure is composed by a contribution which accounts for the interactions between amino acids and a contribution which accounts

for protein-solvent interactions.

- Although the statistical method considers residue-wise probabilities, the decision of the type of contact performed by two amino acids is made by considering the full atomistic configuration of the system.

- The types of contact a couple of amino acids can make are mutually exclusive and include secondary-structure-specific contacts, Van der Waals contacts and also the absence of contacts. The type of contact is checked with a system of priorities: first, contacts proper of secondary structure elements are checked via the DSSP algorithm [KS83]. Then, sidechain-sidechain Van der Waals contacts are searched, by considering the least distance between the heavy atoms of the sidechains of the two residues. If none of these types of contact is found, the two amino acids are labeled as not in contact.

- The solvent interaction is accounted for by checking if each amino acid is exposed to the solvent. This is done by calculating the solvent exposed surface per residue.

The score associated to a structure is the sum of two contributions:

$$E_{\text{BACH}} = pE_{\text{pair}} + E_{\text{sol}} \tag{2.8}$$

where p is a parameter fixed at 0.6, and E_{pair} and E_{sol} are statistical potentials that account for the pairwise contacts and the protein-solvent interactions, respectively [CGL$^+$12].

2.2.2 Pairwise contact term

For each residue (or couple of residues) considered, an appropriate free-energy-like term $\epsilon_{a_i,a_j}^{c(ij)}$ is added to the score:

$$E_{\text{pair}} = \sum_{i<j}^{N} \epsilon_{a_i,a_j}^{c(ij)} \tag{2.9}$$

where N is the total number of residues of the protein being scored, i and j are indexes running on the residues, a_i and a_j are the amino acid type of residue i and j respectively and $c(ij)$ is the type of contact between residues i and

j. The contributions $\epsilon^{c(ij)}_{a_i,a_j}$ are calculated on a training set of native protein structures (see Section 2.2.4) in the following way:

$$\epsilon^{c}_{a,b} = -\log\left(\frac{P(c|ab)}{P(c)}\right) = -\log\left(\frac{\frac{\sum_\alpha n^{c,\alpha}_{ab}}{\sum_\alpha \sum_{c'} n^{c',\alpha}_{ab}}}{\frac{\sum_\alpha \sum_{a'b'} n^{c,\alpha}_{a'b'}}{\sum_\alpha \sum_{a'b'} \sum_{c'} n^{c',\alpha}_{a'b'}}}\right) \tag{2.10}$$

where a and b are indices which run on the 20 types of residues, c runs on the 5 types of contacts and α on the structures of the learning database. The number $n^{c,\alpha}_{ab}$ is symmetric in the permutation of the indexes a and b, as it is constructed as

$$n^{c,\alpha}_{ab} = \frac{\#res^{c,\alpha}_{ab} + \#res^{c,\alpha}_{ba}}{2} \tag{2.11}$$

where $\#res^{c,\alpha}_{ab}$ is the number of residues of type a which make a contact of type c with a residue of type b, observed in structure number α of the learning database.
The five classes of contacts are:

- *Parallel β-sheet* ($c = 0$): this kind of contact is identified by checking the presence of the two hydrogen bonds between the two considered residues using the DSSP [KS83] algorithm. If the hydrogen bonds have energy $E_{hb} < -0.5$ kcal, according to the potential energy function in Ref. [KS83], the bond is formed. This method is used also with the other two secondary structure contact classes.

- *Antiparallel β-sheet* ($c = 1$).

- *α-helix* ($c = 2$).

- Sidechain-sidechain Van der Waals contact ($c = 3$): this type of contact is identified by checking the minimum distance between two atoms belonging respectively to the first and to the second residue considered. If the distance is smaller than 4.5 Å, the bond is formed. By priority, this kind of contact can be formed only if a secondary structure contact between the same two residues is not present.

- Non-contact ($c = 4$): the two residues are considered in non-contact if none of the previous types of contact is formed.

The kind of operation BACH makes to merge the information from all the different structures contained in the learning database is an average. Indeed, in Eq. 2.10 we can see that the sum on the structure index α appears before every occurrence of $n_{ab}^{c,\alpha}$. This implies that BACH extracts the same amount of information by learning the parameters on one conformation or on a thousand replicas of the same conformation. It also implies that, for an optimal learning, the different features observed by BACH must be present in an appropriate proportion in the database. In other words, if the learning database is biased on some particular form of contact, BACH parameters will be biased as well. This "bias transfer" is in fact desirable, as it gives a clear way to optimize the parameters: the more reliable the learning set is, the better the parameters will store the appropriate chemical and physical information.

2.2.3 Solvation term

The solvation term is a one-body score term constructed in an analogous way of the pairwise term. Again, for each residue a term is summed to the score:

$$E_{\text{sol}} = \sum_{i}^{N} \lambda_{a_i}^{e(i)} \tag{2.12}$$

where the parameters $\lambda_{a_i}^{e(i)}$ are calculated on the same training set used for the parameters of the pairwise term, in the same fashion:

$$\lambda_a^e = -\log\left(\frac{P(e|a)}{P(e)}\right) = -\log\left(\frac{\frac{\sum_\alpha n_a^{e,\alpha}}{\sum_\alpha \sum_{e'} n_a^{e',\alpha}}}{\frac{\sum_\alpha \sum_{a'} n_{a'}^{e,\alpha}}{\sum_\alpha \sum_{a'} \sum_{e'} n_{a'}^{e',\alpha}}}\right) \tag{2.13}$$

Here, the environmental classes are only two. The residue can be

- *Exposed* to the solvent ($e = 0$)
- *Buried* in the bulk of the protein ($e = 1$)

The exposure of the residue is inferred by the calculation of its solvent exposed surface area. This quantity is computed by the SURF tool [VBW94] of the Visual Molecular Dynamics (VMD) program [HDS96]. The tool calculates the molecular surface area (MSA) of a protein by building a triangulation of its

surface with the help of a solvent probe. The method will be described in detail in Chapter 3. Here, we only report that the value chosen for the atom and probe radii is 1.8 Å, and that a residue exposed surface composed of 10 or more triangles identifies the residue as exposed.

2.2.4 Training set

In order to learn the two sets of parameters, the TOP500 database [LDA+03] was used. This collection includes 500 non-redundant single domain protein conformations extracted from monomeric and multimeric PDB protein structures. Their size varies between 25 and 840 amino acids. The structures have been solved with resolution better than 1.8 Å by X-ray crystallography. The conformations contain unstructured parts. It was checked that the information extracted from this database is highly correlated with the one extracted from the CATH database [OMJ+97].

2.2.5 Assessing the performance of BACH and other state-of-the-art scoring functions

In order to assess the performance of a scoring function for protein folding, one can try to discriminate the native state of a protein among a large set of wrong conformations. These sets are usually called decoy sets, and can contain conformations produced by a single algorithm or predictions coming from different sources. Once every conformation in the decoy set (native state included) has been scored, the scores are sorted from the lowest to the highest, or viceversa, depending on the definition of the score. The position of the native state in this rank determines the performance of the scoring function: if the native state is the first one in the rank, the scoring function discriminates it correctly. The lower the rank of the native conformation is, the better.

Thus, for each decoy set considered, a normalized rank is produced by dividing the position of the native state in the rank by the number of structures scored. Then, normalized ranks are sorted from the lowest to the highest. In this way, a line will be produced in which the first point is the normalized rank on the decoy set in which the scoring function performed better, and the last on the decoy set in which the scoring function performed worse. Fig. 2.1 reports such graph. In Ref. [CGL+12], 33 decoy sets from CASP 8/9 [MFK+09] were

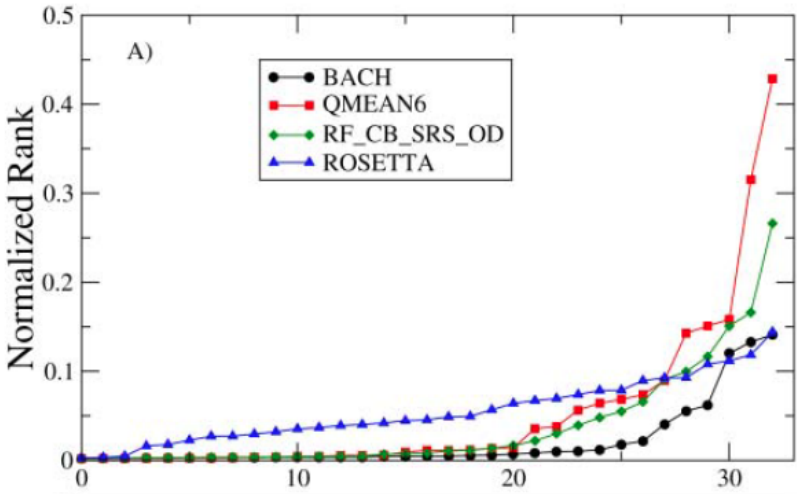

Figure 2.1: Normalized ranks sorted for the decoy sets in CASP 8-9, and calculated for the BACH, QMEAN6, RF_CB_SRS_OD and Rosetta scoring functions (from Ref. [CGL+12]).

considered. Such decoy sets were chosen because they were the most difficult available ones for the discrimination of native structures [CGL+12]. The decoys and the native state have all the same sequence.
The state-of-the-art scoring functions used to compare BACH performances are QMEAN6 [BTS08], RF_CB_SRS_OD [RF10] and Rosetta [SRK+99]. BACH obtains the best performance of them all.
Every time we will need to assess the performances of BACH for problems of protein folding (see Chapter 4), we will refer to this analysis.

2.3 Scoring functions for protein-protein interaction

Parallel to the development of scoring functions for protein folding problems is the one of scoring function for protein-protein interaction problems. A large variety of knowledge-based potentials are available, some of which will be described in this section. Nonetheless, there are very few which can tackle both PF and PPI problems. One of the scopes of this thesis will be deriving such a scoring function, starting from the scoring function BACH presented in Section 2.2.
In order to summarize the most used features of the state-of-the-art approaches to PPI, we will now present six algorithms which will be used in Chapters 4

and 5 for quality assessment. Although all six methods provide a complete procedure for docking, only ZDOCK will be used for that scope (see Chapter 5), while for the other five only the scoring function will be considered.

2.3.1 Rosetta

We already described the statistical method behind Rosetta in Section 2.1.2. The Rosetta algorithm is used here for protein-protein interaction problems with the set of weights "score12". This set of weights is the all-atom, "general purpose" one and is not optimized for interface prediction. It is, nonetheless, one of the most used by the community.

2.3.2 PIE/PISA

PIE [DE10] and PISA [VRE13] are recent scoring functions devised by Elber and co-workers. They are founded on a residue-based and an all-atom description, respectively. The approach is based on Sippl's formalism:

$$\begin{aligned} U &= \sum_{i>j} W_{a_i,a_j}(r_{ij}) \\ W_{a_i,a_j}(r_{ij}) &= -\log\left(\frac{P_{a_i a_j}(r_{ij})}{P^{\text{ref}}_{a_i a_j}(r_{ij})}\right) \end{aligned} \tag{2.14}$$

where i and j are indexes running on the amino acids of the analyzed structure, a_i and a_j are the types of amino acids and $P^{\text{ref}}_{a_i a_j}(r_{ij}) = P_{a_i} P_{a_j} P(r)$. The learning method is based on solving a very large ($\sim 10^{13}$) set of inequalities with a maximum margin method [YJEP07]. Specifically, the set of parameters $\mathfrak{P}$ must respect a large number of inequalities of the form

$$U(X_n;\mathfrak{P}) - U(X_d;\mathfrak{P}) > 1 - \frac{\eta_{n,d}}{\Delta_{n,d}} \tag{2.15}$$

$$(\mathfrak{P},\eta) = \min\left(|\mathfrak{P}|^2 + C\sum \eta_{n,d}\right) \tag{2.16}$$

where X_d are the coordinates of a decoy, X_n are the coordinates of the native, $\Delta_{n,d}$ is the iRMSD between the two conformations and $\eta_{n,d}$ is the number of violated inequalities in the set composed by the native state and all its decoys. The parameters $\mathfrak{P}$ are then found by minimizing the score defined in the right hand side of Eq. 2.16. This is accomplished via the primal-dual recursive method.

PISA [VRE13] is the atom-wise counterpart of PIE, and was added in a second moment in order to enhance the performance of the re-scoring function in the PIEDOCK algorithm. In the article presenting the new potential, three methods to combine PIE and PISA scores are explored. We will use the first one, which multiplies the two values given by the scoring functions.

2.3.3 IRAD

IRAD [VHW11] is the latest version of the well-known algorithm ZRANK [PW07]. The original scoring function was introduced to rank the structures predicted by the docking program ZDOCK [PHW11]. It considers a linear combination of atom-wise energy terms weighted by a set of optimized parameters:

$$\begin{aligned} E = {} & W_1 \cdot VdW_{\text{attr}} + W_2 \cdot VdW_{\text{rep}} + W_3 \cdot Q_{\text{attr, sr}} + W_4 \cdot Q_{\text{attr, lr}} \\ & + W_5 \cdot Q_{\text{rep, sr}} + W_6 \cdot Q_{\text{rep, lr}} + W_7 \cdot \text{ACE} \end{aligned} \tag{2.17}$$

The terms include Van der Waals attractive and repulsive terms, electrostatic attractive and repulsive, long range and short range terms and ACE, a statistical contact potential derived from monomeric protein structures and taken from the ITASSER algorithm [Zha08]. The parameters W_1-W_7 are optimized by minimizing the rank obtained by scoring 93 decoy sets of complexes extracted from the ZLAB Benchmark 3.0 protein-protein database [HPM$^+$08], with 54000 decoys each created by the program ZDOCK [PHW11].
IRAD also includes four residue-based potentials, three of which for protein-protein interaction and one for protein folding. Its performances are shown to be better than those of ZDOCK and ZRANK, which already attained a good prediction power on many CAPRI decoy sets.

2.3.4 HADDOCK

HADDOCK is a well-known docking program [DBB03] which always participated to CAPRI competitions. The docking procedure is divided into different stages in which various degrees of refinement are applied. The scoring has been optimized for each stage of the protocol: we will consider the score of the last refining stage. The energy estimate in the last stage is computed as a weighted

sum of different physical terms:

$$E = 1.0E_{\text{VdW}} + 0.2E_{\text{elec}} + 0.1E_{\text{AIR}} + 1.0E_{\text{desolv}} \tag{2.18}$$

The first two terms describe the Van der Waals and electrostatic contributions, respectively. The term E_{AIR} considers a set of *ambiguous interaction restraints* which use specific structural data to guide the potential towards a more restricted set of plausible conformations. The last term accounts for the desolvation energy relative to the interface area.
The scoring function is not available as a standalone program, but the authors use it in CAPRI scoring competitions, obtaining excellent results [JHM$^+$03].

2.3.5 FireDock

Another docking program whose scoring function is used in CAPRI scoring competitions is PatchDock [DNW02]. The corresponding internal scoring function is called FireDock [ANW07]. FireDock is a physical scoring function based on binding free energy estimations, solvation (using the ACE algorithm from ITASSER [Zha08]), electrostatics, Van der Waals, hydrogen bonds, rotamer torsion energies, π-stackings, aliphatic interactions, and the degree of exposure of the residues. The weights of the corresponding terms are optimized through a linear programming - support vector machine approach. The scoring function was tested also as a refinement algorithm for docking programs other than PatchDock, notably ZDOCK and RosettaDock. In both cases, FireDock succeeds in refining the results of the docking programs better than their internal scoring functions do. FireDock has also been tested in CAPRI competitions, with positive results.

2.4 ZDOCK rigid docking algorithm

In Chapter 5 we will need to generate a large set of conformations of a dimer in order to assess the performance of BACH and other scoring functions in discriminating the near-native conformations of a protein-protein complex. For this scope, we decided to rely on a single rigid docking algorithm, in order to eliminate biases due to the use of different algorithms and to internal scoring functions employed in flexible docking procedures. We thus choose ZDOCK, [PHW11] a rigid body docking algorithm based on Fast Fourier Transform

(FFT) calculations. It takes into account shape complementarity, desolvation and electrostatics and should be used in combination with ZRANK [PW07], an energy minimization algorithm for refining and reranking ZDOCK results. The separation between the rigid docking algorithm and the refinement methods allows us to employ only the former. The rigid docking algorithm leaves the receptor fixed and rotates the ligand around the receptor. The ligand is moved by a constant amount of degrees in the direction of one of the three Euler coordinates (θ, ϕ, ψ). Once the rotation has been performed, ZDOCK makes use of FFT calculation to account for translations: it shifts the two subunits by uniform steps along the axis passing from the two centers of mass and only retains the configuration which corresponds to the maximum value of the internal scoring function of the algorithm. The scoring function is based on three weighted terms which account for shape complementarity, desolvation and electrostatics respectively. This results in pairing the configurations with coordinates (θ', ϕ', ψ') and $(-\theta', -\phi', \psi')$: only one of them will be retained, while the other will be discarded. In particular, the finest sampling that ZDOCK can attain is by generating a configuration for each rotation of 6 degrees. Thus, $180 \times 360 \times 360/6^3 = 108000$ configurations are expected, but since opposite configurations are paired, only 54000 configurations are obtained. This is the only source of unevenness that the algorithm introduces.

Chapter 3

Estimating the solvation propensity

Among the forces that guide the folding and the protein interaction processes, solvent interactions are thought to be the most important [Kau59, Cha05, Bal14]. To trigger folding, hydrophobic side chains localize into the core of the protein, surrounded by the polar and charged functional groups, which interact more favorably with water molecules. Despite its central role for the stability of proteins and protein complexes, modeling the interaction with the solvent is still considered a hard task. Treating the solvent classically, simulating a large protein in a box of explicit solvent remains still very expensive. This hinders the use of explicit solvation for massive screening. For example, in the contexts described in Chapter 2, we need a method to estimate the solvent interaction with a protein in a time comparable with the one needed to evaluate the other energy terms. Going down the ladder of coarser and coarser approximations, we find explicit coarse-grained water models, which groups small numbers of water molecules (usually four) into pseudosolvent particles [MRY$^+$07]. If this is still too slow one has to pass to a continuum representation. Models based on the Poisson-Boltzmann equation are thought to be the most accurate of this class [HN95]. Nonetheless, even numerical solutions of that equation can prove to be too slow to meet the speed requirements of a scoring function. Moreover, the Poisson-Boltzmann equation by itself does not model the entropic contribution to the solvation energy, which many times is dominant [Cha05]. A faster class of models is instead based on the generalized Born equation with surface area correction [CC84, DB02]. Alternatively, very fast approaches based on interatomic distances or on the exposure of particles to the solvent

have been considered [Cho74]. The solvation term contained in BACH and introduced in Chapter 2 belongs to this last class: each residue is considered exposed to the solvent or buried in the protein depending on the relative value of the exposed surface area. Then, an "energy" is assigned to the residue taking into account this attribute and the chemical nature of its side chain. In this chapter we will describe the implementation in BACH of a suitably modified LCPO method [WSS99] to calculate the exposed surface area of each residue. The method is faster than the one exploited in the original implementation and allows computing the derivative of the exposed surface area with respect to the coordinates. In order to make the algorithm faster, we will limit the number of the LCPO parameters to two: the probe radius and the threshold value to determine if a residue must be considered buried or exposed to the solvent. We will show that with an appropriate optimization of these two parameters, the method can still attain an excellent precision. In order to assess the performances of our implementation of the LCPO method, we will compare its results with the ones provided by two algorithms for the determination of the solvent exposed surface area: SURF [VBW94], which is the algorithm previously used in BACH, and GETAREA [FB97], which is commonly used as a reference [KTSW11].

3.1 Solvent accessible surface and solvent excluded surface

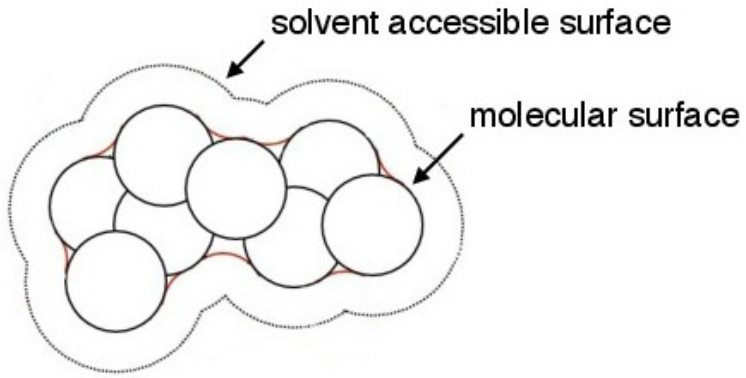

Figure 3.1: Solvent accessible surface (SAS) and molecular surface (MS).

In order to calculate the contribution of each residue or atom to the solvation free energy term, a method to calculate their exposed surface is needed. However, there is a certain amount of ambiguity in defining this concept. There are indeed two commonly used definitions of exposed surface: the *solvent accessible surface* (SAS) and the *solvent excluded surface*, or *molecular surface* (MS). Both are represented in Fig. 3.1. The SAS corresponds to the dotted line, and is defined as the locus of points the centre of a probe sphere visits when it moves in contact with the solid representing the molecule. The MS is composed by all the points of the surface of the molecule with which the probe sphere can be in contact. In the regions in which the probe sphere fails to touch the surface, the MS is the surface of the probe. In Fig 3.1 these patches are represented in red. One can also see the MS as the boundary of the volume not accessible by the probe sphere.
Although their geometric properties differ, the estimations of the exposed surface given by considering the definitions of SAS and MS are closely related, and often applied with no distinction. We will see that, in specific situations, the two measures can actually provide different results.

3.1.1 Methods to calculate the SASA

The solvent accessible surface area was introduced for the first time by Lee and Richards in 1971 [LR71], who also provided a first procedure to estimate it. Since then, many exact and approximate methods aimed at computing it were developed: analytical and numerical approximations were proposed, as well as exact analytical expressions. The Shrake-Rupley algorithm [SR73], one of the most widely used numerical approaches, is also one of the earliest. For each atom in the molecule the algorithm draws a fixed number of points equidistant from the sphere representing the atom. The distance from the atom is typically the radius of the solvent molecule one wants to use, which in the case of water is usually set to 1.4 Å. Each point is checked against the distance from the surface of every neighboring atoms to determine if it is buried or accessible to the solvent. The fraction of the points relative to one atom that is considered accessible by the solvent determines the SASA of that atom. The Shrake-Rupley algorithm and other numerical methods can be simply implemented, but the calculation is usually lengthy and further decreases its efficiency when also the derivative of the surface area with respect to the coordinates must be computed.

Instead of relying on approximated strategies one can carry out the analytical calculation of the surface area of a solid resulting from the union of many spheres. Among the exact methods, we recall the calculations through integration using the Gauss-Bonnet theorem [Ric84, FB97] and more recently via the construction of power diagrams [KTSW11]. GETAREA [FB97], a method based on the Gauss-Bonnet theorem, proved to be very reliable over the past decades. We will thus use it as a reference to benchmark our approach.
In general, exact analytical methods are very powerful: their estimates have uncertainties of tenths of angstroms on the surface of whole proteins and allow calculating the derivatives of the surface area with respect to atom coordinates. However, they are often rather computationally expensive, and are usually complex to implement. Approximated methods compensate their coarser calculations with an increased overall speed and an ease in calculating the derivatives of the area with respect to the coordinates of the atoms. Among them, methods that reduce the analytical summation of intersections to a sum of pairwise contributions have been explored for decades [Ric84]. A recent method based on this strategy is the Linear Combination of Pairwise Overlaps (LCPO) [WSS99], which we will describe more in detail in Section 3.1.4.

3.1.2 Methods to calculate the MSA

The molecular surface conveys a more complete information about the shape of the molecule, but unlike the SASA it can be calculated only using approximations [Ric77]. A well-known algorithm that estimates the molecular surface area is the rolling-probe algorithm [Con83]. Another approach employed in many recent algorithms is based on considering the cloud of points given by the centers of the spheres that compose the solid and calculating its α-complex, first defined by Edelsbrunner, Kirkpatrick and Seidel in 1983 [EKS83]. The α-complex consists in a piecewise linear surface composed by triangles. Operatively, a triangle is drawn for all the triplets of points that sit on the boundary of a probe sphere of radius $\alpha^{1/2}$ that contains no other point. The computational cost of calculating an α-complex can be reduced significantly. For this reason many algorithms first build the α-complex of the cloud of points and then arranges surface patches on it in order to recover the molecular surface. It must be stressed that in all these methods the radius of the probe sphere is a crucial parameter and greatly affects the outcome. Among the methods which use the α-complex to estimate the molecular surface area, we will consider

SURF [VBW94], an algorithm implemented in Visual Molecular Dynamics (VMD) [HDS96] and used in the original BACH method for the estimation of the solvation environmental class of the residues of a protein.

3.1.3 VMD SURF tool

SURF implements an algorithm based on a mixed approach, in which a preliminary scan by power subgraphs is used in order to decide whether an atom is buried or could be exposed. In a second moment, to each of the exposed atoms is assigned a patch depending on the geometrical arrangement of their neighbors. The patch is triangulated, and the number of vertices of the triangles belonging to one atom allows estimating its molecular surface. Since the patches are generated by an α-complex algorithm, their conformation depends on the radius of the probe sphere, which is a parameter of the model. In the first implementation of BACH we imposed that the radii of the probe sphere and of all atoms be set to 1.8 Å. The radius is thus larger than the typical one used both for the water probe (1.4 Å) and for the atoms in the protein (1.3-1.7 Å). This was made in order to avoid including internal cavities in the calculation of the exposed surface area. The output of SURF is the number of triangle vertices associated to each atom of the protein. These vertices are used in the triangulated representation of the protein surface employed by VMD, and it was calculated that the area associated with each vertex is approximately 0.15 Å^2. However, this value is just for reference, for as we will show the dependence of the molecular surface on the number of triangles is not linear. By summing over all the atoms of a given residue, the number of vertices associated to that residue is obtained. In the original BACH implementation, side chains must have at least 10 vertices (~ 1.5 Å^2) to be considered exposed.

3.1.4 LCPO

SURF is proved to be accurate, but in this approach the derivative of the surface area with respect to the atomic coordinates cannot be calculated. An approach that allows computing derivatives is the LCPO [WSS99] method. This approach approximates the SASA of a solute as a linear combination of the surfaces of its atoms, modeled as spheres of radius r. The working principle is to remove from the sum of the whole surfaces the estimated overlap of the surfaces of nearby atoms. The exposed surface of the atom i is approximated

as follows:

$$A_i = P_1 4\pi r_i^2 + P_2 \sum_{j \in N(i)} A_{ij} + P_3 \sum_{\substack{j,k \in N(i) \\ k \in N(j) \\ k \neq j}} A_{jk} + \\ + P_4 \sum_{j \in N(i)} A_{ij} \left(\sum_{\substack{k \in N(i) \\ k \in N(j) \\ k \neq j}} A_{jk} \right), \tag{3.1}$$

where

$$A_{ij} = 2\pi r_i \left(r_i - \frac{d_{ij}}{2} - \frac{r_i^2 - r_j^2}{2d_{ij}} \right); \tag{3.2}$$

r_i is the radius of atom i, $N(i)$ is the list of atoms that overlap with atom i and d_{ij} is the centre-to-centre distance of atom i and j. In the original work [WSS99] the four parameters P_1-P_4 depend on the hybridization of the atom and on its neighborhood and are estimated by linear regression on a heterogeneous database of analytically calculated cases. Only the elements appearing in amino acids or DNA bases are parametrized. This does not allow using the original LCPO method to calculate the SASA of a generic molecule, e.g. a drug binding to a receptor.

3.2 Scoring solvation by modified LCPO

The original BACH algorithm used the SURF tool of VMD to estimate the exposure of the residues of a protein to the solvent. Although SURF provides an accurate estimate of the solvent exposed surface area, it is rather slow and does not allow to calculate the derivatives with respect to the atomic coordinates. These drawbacks made us abandon the tool in favor of the approximated LCPO [WSS99] method. While SURF calculates the molecular surface area (MSA), LCPO calculates the solvent accessible surface area (SASA). This quantity can be in principle evaluated exactly, for example with the algorithm implemented in GETAREA. LCPO is a faster method which relies on an approximation of the analytical formula. Likewise the exact method, LCPO provides an estimation of the SASA explicitly dependent on the coordinates of the molecules. The main drawback of the method is the dependence on a quite large set of parameters (four for each different element and hybridization), which also limits the range of possible molecules that can be evaluated: indeed, the parameters are only

provided for the species contained in the amino acids and nucleic acids. We thus decided to modify the method by reducing the quantity of parameters to only two. We then optimize these parameters in order to achieve a good performance. During all the procedure, we assess the quality of the exposed surface area estimations both with SURF and with the standard reference algorithm GETAREA.

3.3 Modified-LCPO (mLCPO)

We decided to drastically reduce the number of parameters in the model imposing:

- a constant radius r_i for all the heavy atoms;
- a constant set of four parameters P_1-P_4 for every heavy atom.

Specifically, the parameters for the sp3-carbon bound to three heavy atoms were used, since it is one of the most common atom species throughout the 20 types of amino acids. Since the parameters P_2 and P_3 are negative, the value of A_i can also be negative. Thus, the value estimated by Eq. 3.1 is meaningful only in a relative way. To account for this coarse approximation, we promote the radius of the atoms to a free parameter, and we optimize its value with an iterative procedure. We call this method *modified-LCPO* (mLCPO).

3.3.1 Coherence score

For optimizing the radius we need to devise a meaningful estimator to assess the performance of the mLCPO algorithm. For this scope, we define the *coherence score* as the fraction of residues over the proteins of the TOP500 dataset for which SURF and the mLCPO algorithm agree on the environmental class assignation (exposed or buried).
The coherence score is an indirect estimator: in order to produce an outcome, a threshold between the two environmental classes must be set. The analysis must then be performed by taking into consideration two free parameters rather than only one: the atomic radius and the buried/exposed (*in/out*) threshold. We eventually take the couples of values that maximize the coherence score.

3.3.2 Using GETAREA as reference

In order to benchmark the general consistency of this method, we also repeat the procedure by taking the GETAREA estimation of the exposure instead of the SURF one. We do this in order to compare the mLCPO estimation with another estimation of the same quantity (the SASA) instead of an estimation of a slightly different one (the molecular surface computed by SURF). Moreover, we believe the GETAREA estimation of the exposed surface to be more precise than that of SURF, both because it is based on an exact analytical calculation.

3.4 Results

The method mLCPO described in Section 3.2 allows estimating very quickly the SASA. Its reliability is tested against two algorithms for the calculation of the exposed surface area of a protein, SURF and GETAREA. To be able to apply the coherence score defined in Section 3.2, we have to define a suitable buried/exposed (in/out) threshold for these two approaches. For the method SURF, we rely on the choice made in [CGL$^+$12] to consider a residue as exposed if more than 10 vertices are found. For the method GETAREA, we choose a threshold by looking at the distribution of residue-wise SASA values calculated by the two algorithms for the residues of the proteins of the TOP500 database. The frequency distributions for both SURF and GETAREA estimations of the exposed surface areas of the residues of all the proteins in the TOP500 database are reported in Fig. 3.2. Both distributions present a peak at zero value corresponding to the deeply buried residues. However, the shape of the distributions for small but positive values differs significantly: SURF does not assign small nonzero values, while GETAREA has a more continuous transition. This implies that while for low threshold values SURF is not sensitive to small variations, GETAREA is. This confirms the need of properly tuning the threshold parameter in order to have consistent estimates. By setting the threshold for GETAREA such that the number of residues considered buried is the same than the one obtained with SURF we find a value of 1.9 Å.

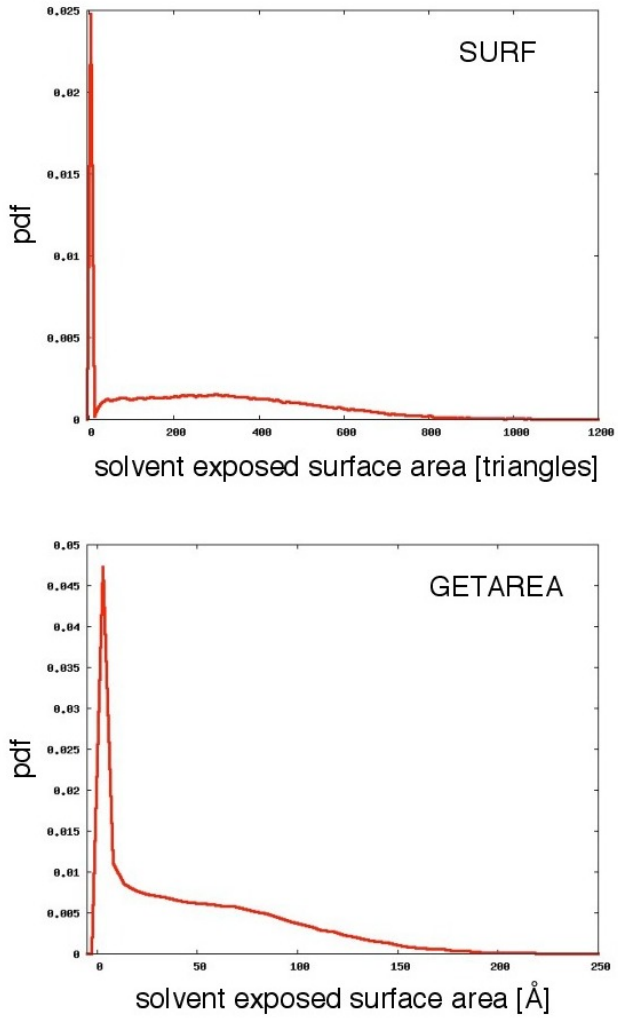

Figure 3.2: Probability distribution of SURF and GETAREA estimates of the MSA and SASA of each residue in the protein of the TOP500 database.

3.4.1 Coherence score between different estimates of residue exposure

We now benchmark the reliability of mLCPO with respect to the recognition of residues that are exposed to the solvent. We accomplish this task by computing the coherence score defined in Section 3.2. We first compare SURF and GETAREA in order to assess the reliability of the two reference methods. We find a coherence score of 0.96, indicating a very good agreement: the two functions disagree on the environmental class for roughly 1 out of 25 residues. In Fig. 3.3 we plot the coherence of mLCPO with respect to SURF as a function of the probe radius, for different values of the threshold. The two estimates are in good agreement, and the maximum value of the coherence

score is 0.84, meaning that less than 1 residue out of 5 is assigned to the wrong environmental class according to SURF.

The graph shows that the maximum value has a one-dimensional quasi-

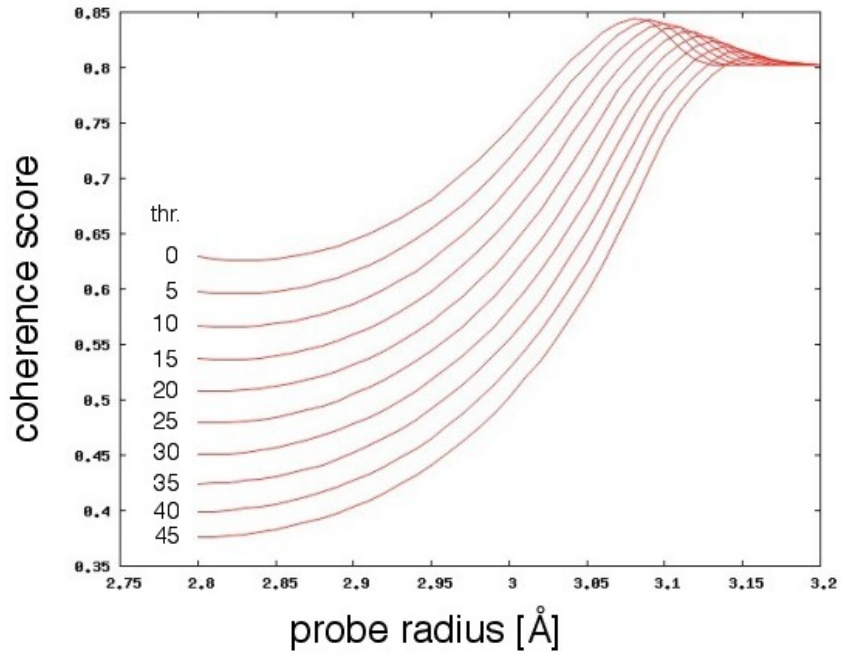

Figure 3.3: Coherence score of the mLCPO method with respect to SURF. The coherence score is defined as the fraction of residues of the proteins of the TOP500 dataset for which SURF and the mLCPO algorithm agree on the environmental class assignation (exposed or buried). It is a function of the mLCPO probe radius and the in/out threshold. Here, the coherence score is shown as a function of the probe radius, for different fixed values of the in/out threshold.

degeneracy, indeed the maximum value decreases only slowly as the threshold value increases. Moreover, for low thresholds the maximum is broader. Lastly, the coherence score is higher on the left of the peak area at low threshold values. For these values of the threshold the dependence on the radius is thus less marked. All these observations make us decide to consider a threshold value of 0 and a radius of 3.08, for which the coherence score is maximum. By repeating the analysis with respect to GETAREA we find that the two optimal parameters are almost undistinguishable (Fig. 3.4).

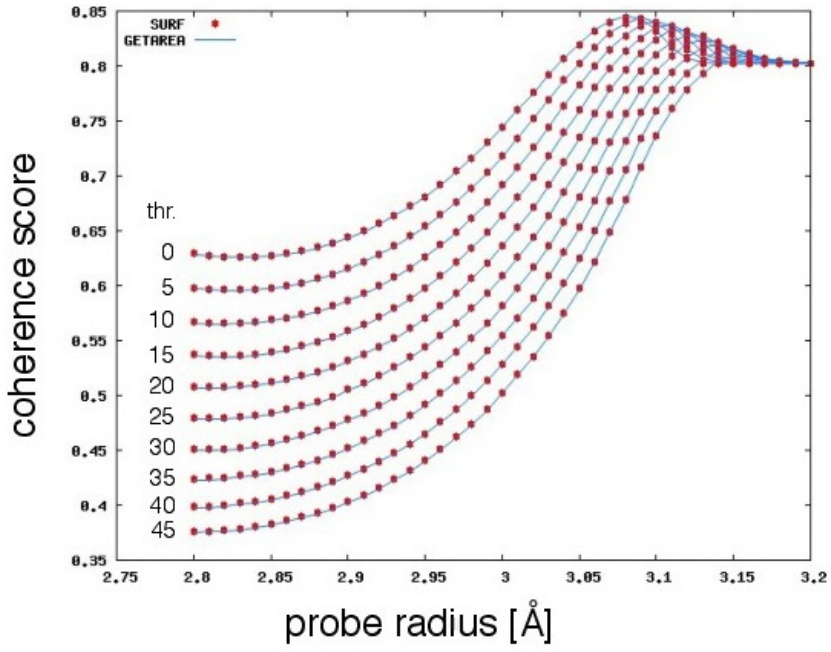

Figure 3.4: Coherence score of the mLCPO method with respect to SURF (red dots) and GETAREA (blue lines). See caption of Fig. 3.3. The scores with respect to the two reference methods are almost undistinguishable.

3.4.2 Optimizing the performance of mLCPO in protein structure prediction

To validate the choice of the threshold we test the performance of the solvation part of BACH on a selection of 10 decoy sets from the ones belonging to CASP rounds 8 and 9 [CKT09]. As explained in Chapter 2, we make use of the normalized ranking as the preferred estimator to compare the performances of the scoring functions. In this framework, we extend the use of this estimator to evaluate the performances of the BACH solvation part only. Throughout this section, the score of BACH will thus coincide just with E_{sol} defined in Section 2.2.3. To perform the optimization by ranking, we choose the 10 decoy sets reported in Table 3.1 from the ones in CASP 8-9 and evaluate the performance of BACH by looking at the rank of the native conformation.

As it was done for the coherence score, we scan the average rank landscape by varying the threshold and radius parameters. In Fig. 3.5 we show the average solvation rank and the SURF coherence score as a function of the probe radius, fixing the threshold at three representative values: $thr = 0$ (first panel), $thr = 20$ (second panel), $thr = 40$ (third panel). The trend of the

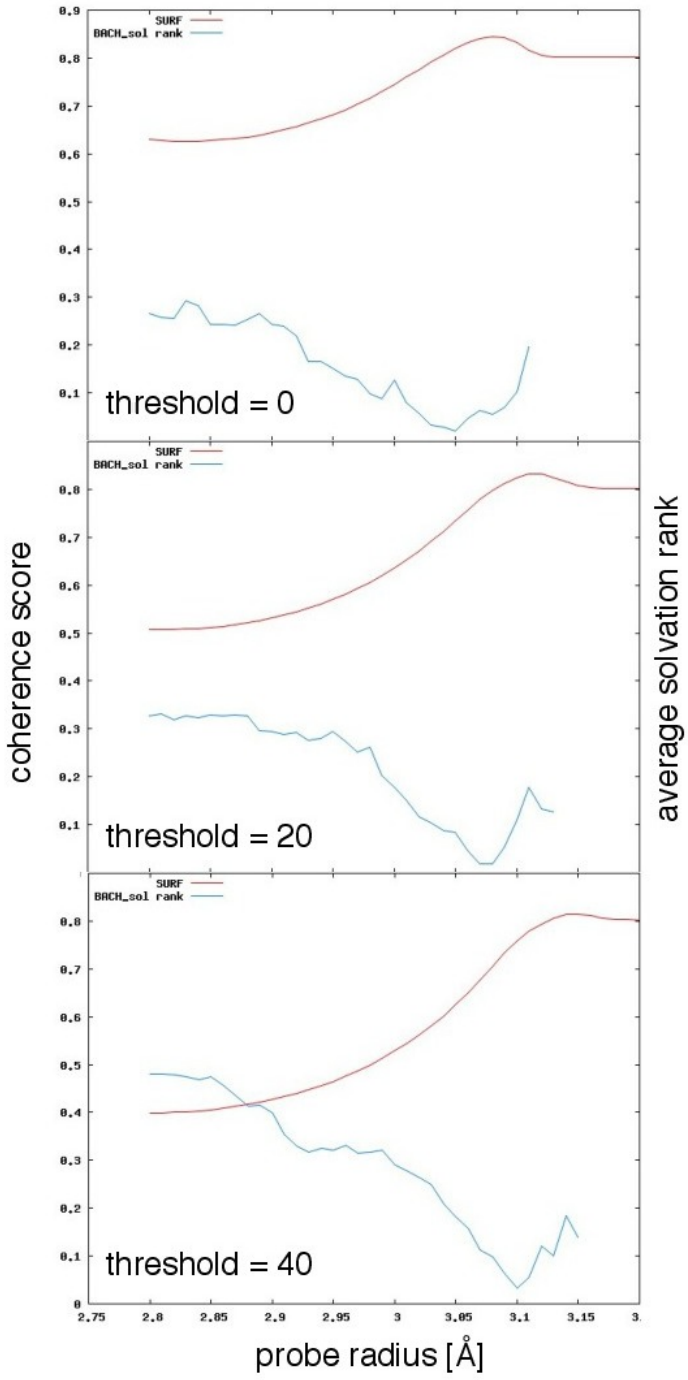

Figure 3.5: Coherence score and average solvation rank as a function of the probe radius for three fixed values of the in/out threshold. For both quality measures, a shift towards higher values of the optimal probe radius is visible for increasing threshold values.

CASP code	PDB code
T0388	3cyn
T0397	3d4r
T0415	3d6w
T0425	3czx
T0427	3d3y
T0432	3dai
T0433	37dl
T0437	2k3i
T0440	3dcp
T0445	3dao

Table 3.1: 10 CASP decoy sets used to calculate the average solvation rank.

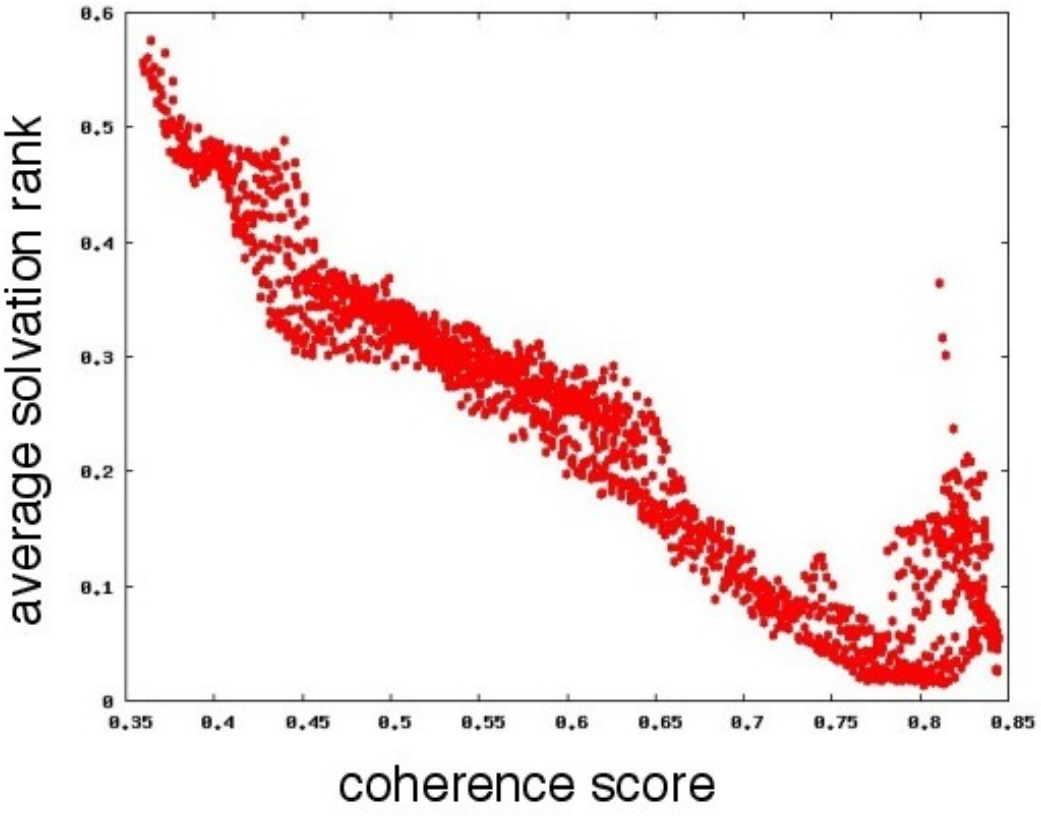

Figure 3.6: Correlation plot between the average solvation rank and the coherence score, for each pair of probe radius and threshold values.

two estimators is very similar, as highlighted in Fig. 3.6, where the correlation between the two estimators is reported for each combination of parameters considered. We also observe a shift between the peaks of the two estimators. Nonetheless, for the case in which $thr = 0$ the coherence score peak selects a value of the probe radius for which the average solvation ranking is below 0.1, and the fourth smallest value.

3.4.3 Comparison of the residue-wise SASA estimates

Now that we fixed the radius parameter we can look at the distribution of the residue-wise SASAs of the TOP500 proteins as estimated by mLCPO.
In Fig. 3.7 we report the mLCPO distribution and the GETAREA distribution

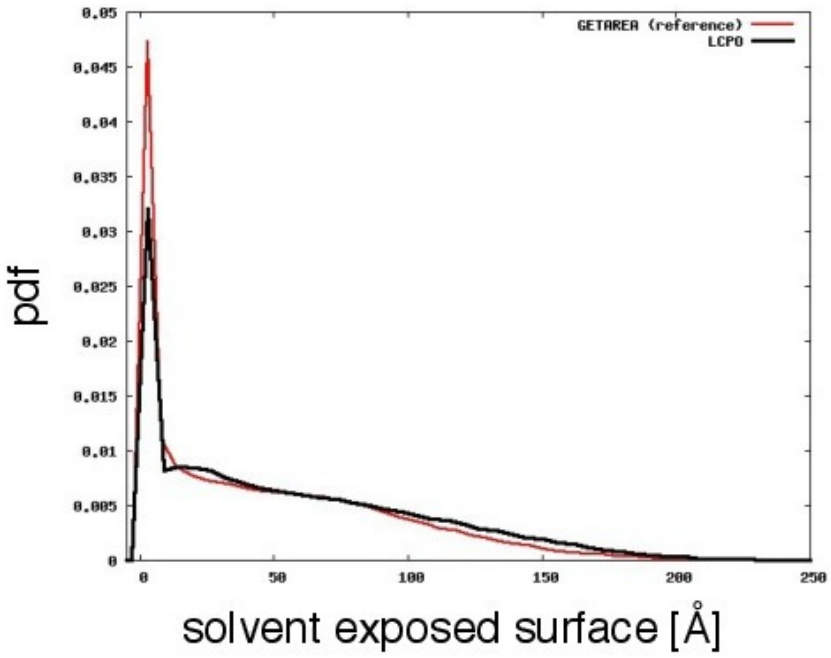

Figure 3.7: Frequency distributions of the SASAs of each residue belonging to the TOP500 database proteins. The distribution for mLCPO is obtained by equating all negative value to zero. Even with the coarse approximations introduced in mLCPO, there is a marked similarity in the shape of the two distributions.

for reference. Since the mLCPO algorithm produces a significant quantity of negative SASA values, we replaced them with zero. We can see that the two distributions are similar, thus confirming that the approximations introduced in the mLCPO method do not change significantly the quality of the estimates.

3.5 Assessing the quality of the solvation rank

As a last but important step, we benchmark the mLCPO-based functional form of E_{solv} by the same procedure followed in the article by Cossio and co-workers [CGL$^+$12] and reported in Chapter 2. For all the 33 CASP 8/9 decoy sets,

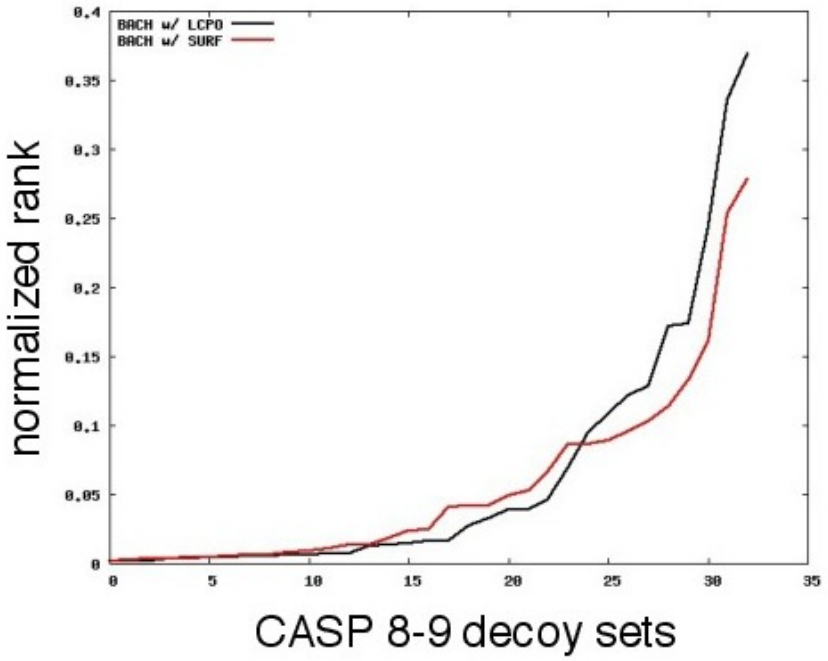

Figure 3.8: Normalized rank calculated on 33 CASP 8-9 decoy sets for the two versions of BACH: with the SURF algorithm and with the mLCPO algorithm. We can see a slight increase in the performance of BACH with mLCPO with respect to the original method.

in Fig. 3.8 the solvation rankings of BACH with SURF and with LCPO are shown. We can see that the new solvation term is slightly more accurate in predicting the native structure on most of the decoy sets, but also that its errors are slightly more relevant when the prediction becomes poor.

3.6 Discussion

In this chapter we implemented and tested mLCPO, a modified version of the LCPO method [WSS99] to calculate the SASA of the residues of a protein. The ability of choosing the correct residue-wise environmental class (buried or exposed) is tested against two reference methods: SURF [VBW94], which is the method previously employed in the molecular surface area calculation in BACH,

and GETAREA [FB97], an exact analytical method for SASA calculation. The comparison is performed in order to choose a suitable value for the two free parameters of the model: the buried/exposed threshold and the radius of the atoms, which is kept constant for every element and includes the radius of the water probe.

The optimization process via the coherence score is thus equivalent to a maximization process in a two-dimensional manifold. In more than one dimension, this kind of problem may have degeneracies if two parameters are not independent. Since the in/out threshold depends on the size of the spheres (and thus on their radius), a degeneracy is likely to show. Indeed, Fig. 3.3 presents such a situation, and thus proves the relation between the optimal threshold and the optimal radius. The degeneracy is not exact as one observes a slight increase of the maximum value of the coherence score as the threshold value decreases. This and the shape of the curve at fixed threshold makes us choose as optimal threshold and radius the values 0 and 3.08, respectively. These values do not differ significantly from the ones reported in [SZC$^+$13].

The coherence score as defined in this chapter is a biased estimator. Indeed, according to SURF (or GETAREA) the TOP500 database is not composed of an even ratio of buried and exposed residues, rather of 20% buried and 80% exposed residues, because of the strict definition of the buried class imposed by the low threshold values. In this case, a scoring function that assigns to every residue the "exposed" environmental class obtains a coherence score of 0.8. This is indeed visible in the saturation value of the graph in Fig. 3.3 and 3.5. A high value of saturation implies a reduction in the robustness of the method. Indeed, the optimal value of the radius is just 0.04 Å away from the saturation point. To remove this bias one can follow two different approaches: changing the in/out threshold of the SURF and GETAREA methods or modifying the database in order to have an equal amount of residues considered buried and exposed by one of the two functions.
By changing the threshold value of SURF we would disrupt the optimization procedure of previous studies [CGL$^+$12] and we would go against intuition, since a natural in/out threshold for SURF is clearly visible in Fig. 3.2. Thus, we decided to consider a subset of the TOP500 database: for each protein, we consider the largest possible amount of residues for which a one-to-one

ratio between exposed and buried can be attained. For example, in a protein containing n exposed residues and m buried residues (according to SURF), and $m < n$, we will consider $2m$ residues: all the buried ones and a subset of m exposed residues chosen randomly among the n total.

In Fig. 3.9 is reported the coherence score of LCPO with SURF when

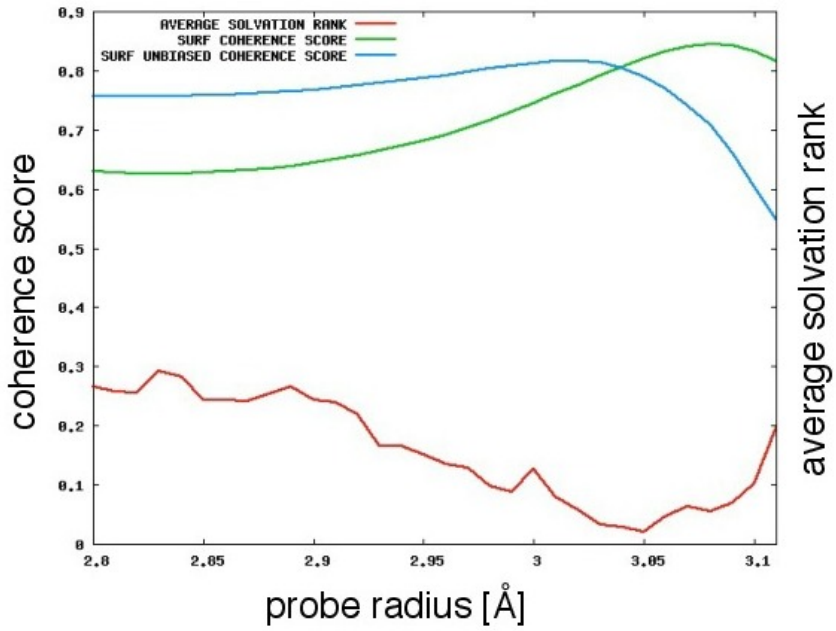

Figure 3.9: Coherence score (green line) and average solvation rank (red line) as a function of the probe radius with the threshold fixed at 0. The unbiased coherence score estimator is represented by the blue line. The coherence score is always calculated as described in Section 3.2, but the unbiased estimator considers a subset of TOP500 in which according to SURF there are as much residues in the exposed environmental class as there are in the buried environmental class.

an equal number of buried and exposed residues is considered. We see that the minimum is more robust, and the optimal value for the radius shifted from $r_{opt} = 3.08$ Å to $r_{opt} = 3.02$ Å. The difference from the optimal probe radius estimated in Section 3.4.2 (3.05 Å) is still small (0.03 Å), but the value is now further from the critical value that corresponds to the saturation of the estimator. The maximum value of the coherence score decreases slightly, from 0.84 to 0.81. We thus conclude that the new definition of the coherence score is preferable.

The average solvation ranking estimator shows to be more noisy than the coherence score. It is however a precious indication: as we mentioned above, a

shift of the optimal radius value is visible independently of the estimator used for the coherence score. Remarkably, for the selected parameters the native state is found on average in the top 10% of the score of the solvation part only.

The frequency distribution of the residue-wise solvent exposed surface area values of LCPO is similar to that obtained with GETAREA, while it differs from the one obtained with SURF. This is expected: likewise GETAREA, mLCPO calculates the SASA, while SURF calculates the MSA. The peculiar qualities of the SURF distribution are as well possibly given by the mixed method it employs: the residues that are assigned with a null area are selected separately by considering a power graph based on a Voronoi tessellation, then the remaining residues are assigned with a non-zero area depending on the number of triangles their accessible surface is divided into [VBW94]. This could generate the separation of the distribution into a delta at zero and a wide peak at large positive values. This separation is not present in the mLCPO and GETAREA methods, in which instead the peak overlaps significantly with the rest of the distribution. The properties of the distribution of SASA values make BACH more sensitive to changes in the definition of the threshold value. This is however a positive fact, since we can choose the threshold value as one of the two free parameters to optimize the model.

Our coarse implementation of the mLCPO method is shown to have a comparable performance to the previous SURF-based algorithm. The rank of the solvation part of BACH score on the 33 CASP 8-9 decoy sets reported in Fig. 3.8 is very similar to the one of the previous version. A little improvement is noticeable for most of the decoy sets. This further validates our approach and definitely allows us to opt for the mLCPO method, which gives us the possibility to implement the derivatives with respect to the SASA. This is an important step towards a version of BACH in which it can be used as an effective potential in enhanced sampling molecular dynamics simulations.

Chapter 4

Extending BACH to protein-protein interaction problems

In Chapters 2 and 3 we presented a scoring function for quality assessment of structures in protein folding (PF) problems. This chapter will be dedicated to extending the scoring function's ability to discriminate the native pose among a set of decoys of protein-protein interaction (PPI). Many knowledge-based scoring functions for PPI already exist [VRE13, VHW11, GMW$^+$03, DBB03, ANW07], but up to our knowledge they are always built specifically for that problem: their parameters are trained on sets of complexes or interfaces. We instead want to develop a scoring function able to discriminate the native pose both for protein folding and for protein-protein interaction problems, without changing the set of parameters.

The reason to pair the quality assessment of PF and PPI processes comes from the observation that, at molecular level, the two kinds of interactions should be indistinguishable: the same sets of atoms produce the same kind of forces (electrostatic, Van der Waals, etc.). However, statistical mechanical considerations reveal substantial differences that could play against this intuitive hypothesis. Indeed, in the two problems enthalpy and entropy differ in magnitude, if not in quality. The enthalpy of protein folding is dominated by the creation of hydrogen bonds in the backbone of the protein as well as by electrostatics and Van der Waals interactions between the residues [Bal07]. Although both interactions are also present in PPI, their relative strength differs significantly: the contribution of electrostatics often dominates, because of the polar and

charged residues present on the surface of the protein [LCCJ99], the hydrogen bonds are less stable [XTN97], and are at times mediated by molecules of water trapped on the interface. There are yet specific classes of heterodimers and homodimers which display interface regions similar to the protein bulk [JT97]. The entropy part is what differs the most: while the conformational entropy of protein folding is dominated by the space spanned by the backbone angles, the entropy in PPI is dominated by the sidechain movements and the loss of rotational and translational degrees of freedom [BS97, Bal07]. The entropy of the solvent also plays a different role, as reported for example in the 2002 review by Scheraga [FS02]. We will show that, despite these important differences documented in the literature, a scoring function built on statistical observations of globular monomers is able to predict the free energy of the different poses of a protein complex. We will base our derivation on a formalism of Information Theory, that we will outline in the next section.

Although these considerations encouraged us to test our scoring function on PPI problems, we were aware that discriminating the native pose of a protein complex presents new issues we were not confronting in the case of single monomers. We identified three of them:

- The different poses of a dimer differ by much less contacts than the different folds of a monomer. Indeed, only the contacts on the interface change, along with a limited number of bulk contacts due to the internal rearrangement of the two subunits. In order to discriminate one pose from another basing on a smaller number of contacts, we had to refine the statistical method on which the scoring function is based.

- Due to the lower quality of the state-of-the-art prediction methods for PPI problems, there is an increased probability to observe steric clashes in the available decoy sets. Clashes are highly disruptive for a contact-based scoring function, because they produce a large quantity of false contacts which are likely to favor unphysical conformations over more correct ones. We thus had to devise a new term to account for steric hindrance.

- The available decoy sets to test the performance of BACH and other state-of-the-art scoring functions often present inhomogeneity with respect to the amino acid sequence of the poses. Many times, poses miss some fragments, often in correspondence with flexible parts. Since the score

is based on extensive properties, it is not fair to compare the score of structures having different amounts of atoms or residues. Thus, we had to devise a method to produce meaningful ranks for the inhomogeneous test sets.

Each of these issues is treated in the next three sections.

4.1 Deriving a scoring function from Information Theory

In this section, an alternative formalism for deriving a scoring function will be proposed. First, we will summarize some useful concepts from Information Theory. Then, we will relate these concepts to the context of BACH statistical method, and we will prove that the formalism can be applied to this special case. The new approach will allow us to improve and extend BACH method in order to devise a scoring function for both protein folding and protein-protein interaction.

4.1.1 The cross-mutual information approach

This section is inspired by the very clear introduction to information and entropy given in [Bia12]. According to the formulation attributed to Shannon, the information that an observed event x gives about a certain system is quantifiable as

$$I(x) = \log \frac{1}{p(x)} \tag{4.1}$$

where $P = p(x)$ is the probability that the event x occurs. It is important to note from the beginning that at the moment of the observation one can ignore the probability of that event to happen. Then, the probability $p(x)$ will be the *supposed* probability to observe such an event. The definition of information given in Eq. 4.1 confers to this quantity some important properties: first, the information given by the occurrence of an event we consider certain is zero ($p(x) = 1 \Rightarrow I(x) = 0$). Second, the information given by two independent events is the sum of the information given singularly by each occurring event ($p(x_1, x_2) = p(x_1)p(x_2) \Rightarrow I(x_1, x_2) = I(x_1) + I(x_2)$).
Let's consider a system that can produce a set of events $X = \{x_1, x_2, ..., x_m\}$. We suppose that these events happen with probability $p(x_i)$ (i.e., they follow the

probability distribution P). After an appropriate amount of N tries, we observe that these events really occur according to the probability distribution P. Then, the information collected after N observations will be $-\sum_{i=1}^{m} Np(x_i)\log p(x_i)$. The entropy of the system is then the average amount of information we collect when we observe an event:

$$S(P) = -\sum_{i=1}^{m} p(x_i)\log p(x_i) = \langle I(P)\rangle \tag{4.2}$$

The second equality allows us to state an even more general definition of this quantity: entropy is the expected value of the information *we* can get out of the probability distribution P. If we take a flat probability distribution, we will have the least hint on what event could occur. Then, when one of those events occurs, we will get the largest amount of information. If we take a probability distribution that associates a probability $p = 1$ to one event and $p = 0$ to all the others, when we will observe that event we will not get any information at all: we already knew the outcome before its occurrence. From this example we understand that the entropy of a probability distribution is *inversely* proportional to the quantity of information the probability distribution already contains.

Most times we do not know *a priori* which is the correct probability distribution of the events we are observing. We thus proceed to evaluate these same quantities using a tentative probability distribution $\tilde{P} = \tilde{p}(x)$, with $\tilde{P} \neq P$. From our point of view $I(x) = \log[1/\tilde{p}(x)]$, and when we try to calculate the entropy, we will instead get the quantity

$$S(P;\tilde{P}) = -\sum_{i=1}^{m} p(x_i)\log \tilde{p}(x_i) = \langle I(\tilde{P})\rangle_P \tag{4.3}$$

which is the expected value (calculated on the actual probability distribution P) of the information of the probability distribution $\tilde{P}$. Indeed, the events will occur according to the actual probability distribution P, yet we are still considering the events as if they were occurring with our tentative probability distribution. The quantity $S(P;\tilde{P})$ is sometimes called *cross entropy* between the distributions P and $\tilde{P}$.
We are now interested to measure how much our model distribution $\tilde{P}$ matches

the actual one. One approach is to use the *Kullback-Leibler divergence*:

$$D_{KL}(P||\tilde{P}) \equiv S(P;\tilde{P}) - S(P) = \left\langle \log\left(\frac{p(x)}{\tilde{p}(x)}\right)\right\rangle_P \tag{4.4}$$

This quantity is not a metric, as it is not symmetric in P and $\tilde{P}$ and it does not satisfy the triangle inequality. However, it can still be used to measure the "distance" between two distributions, because $D_{KL}(P||\tilde{P}) \geq 0$ and $D_{KL}(P||\tilde{P}) = 0 \Leftrightarrow P = \tilde{P}$. Moreover, we stress that

$$D_{KL}(P||\tilde{P}) \geq 0 \Rightarrow S(P;\tilde{P}) \geq S(P) \tag{4.5}$$

from which we understand that the entropy of a probability distribution P is the lower bound of the cross entropy between P and a tentative probability distribution $\tilde{P}$. This suggests that a specific succession of events gives us more information if we are expecting the events to occur with a probability distribution very different from the actual one, and gives us the least quantity of information possible (for that succession of events and for that actual probability of occurrence P) if we are already expecting them to occur according to the right probability distribution. This somehow concurs with our intuition: the more we already know what will happen, the less we will learn when we experience the outcome. The Kullback-Leibler distance can be then interpreted as the additional "surprise" we get when we observe events occurring with a probability density which does not agree with our expectations.

Suppose now that a certain set of events $X = \{x_1, ..., x_m\}$ occurs according to a probability distribution $P_X = p_X(x)$, and consider another set of events $Y = \{y_1, ..., y_k\}$. If the observation of an event y_i tells us something more about the set of events X, the entropy $S(P_{X|y_i})$ of the conditional probability distribution $p_{X|y_i}(x|y_i)$ will be smaller than $S(P_X)$. We identify this reduction in entropy as the *information flux*:

$$I(y_i \to x) \equiv S(P_X) - S(P_{X|y_i}) \tag{4.6}$$

The expectation value over the observations Y of this quantity is called *mutual information*:

$$\langle I(y \to x)\rangle_{P_Y} \equiv S(P_X) - \langle S(P_{X|y})\rangle_{P_Y} \tag{4.7a}$$

$$= S(P_X) + S(P_Y) - S(P_{XY}) \tag{4.7b}$$

$$= \sum_x \sum_y p_{XY}(x,y) \log \left[\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \right] \tag{4.7c}$$

This estimator holds some useful properties:

1. $\langle I(y \to x) \rangle_{P_Y} = \langle I(x \to y) \rangle_{P_X}$
2. $\langle I(y \to x) \rangle_{P_Y} \geq 0$, and $\langle I(y \to x) \rangle_{P_Y} = 0 \Leftrightarrow p_{XY}(x,y) = p_X(x) p_Y(y)$
3. $\langle I(y \to x) \rangle_{P_Y} \leq S(P_X)$ and $\langle I(y \to x) \rangle_{P_Y} \leq S(P_Y)$

The symmetry with respect to the two distributions is particularly notable, since it implies that the average amount of information that P_Y provides when P_X is known is the same than the one that P_X provides when P_Y is known. To stress this property even more, from now on we will refer to the mutual information with the symbol $\mathcal{I}_{XY}(x,y)$. The third property tells us two things: first, that the average amount of information that we can get from P_Y when knowing P_X cannot exceed $S(P_X)$. This was expected, since $S(P_X)$ measures the number of possible states that X can be in, and we cannot learn more than we would learn by choosing one among this whole set of possibilities (i.e. by making one event in X certain and all the others impossible). Second, it tells that the average amount of information that we can get from P_Y when knowing P_X cannot exceed $S(P_Y)$. This is a far less trivial statement, and implies that we cannot learn more on X than what our set of events Y can tell. To state it more clearly, the information we can get about a set of events X (e.g. the states of a physical system) by means of another set of events Y (e.g. the values of an instrument of measure) cannot exceed what the set of events Y can reveal (e.g. the accuracy of the measure). This powerful implication will be exploited in Section 4.1.2.

Lastly, it is important to note that one can also generalize the above mentioned definitions by considering cross entropies instead of normal entropies. A generalized version of the information flux will then be:

$$I_{P;\tilde{P}}(y \to x) \equiv S(P_X; \tilde{P}_X) - S(P_{X|y}; \tilde{P}_{X|y}) \tag{4.8}$$

This *cross information flux* tells us how much information we gain on the occurrence of the events X when we observe another event y, considering

that we are estimating probabilities with inexact probability distributions $\tilde{P}_X$ and $\tilde{P}_{X|y}$ instead that with the actual ones P_X and $P_{X|y}$. The *cross-mutual information* can be defined analogously:

$$\tilde{\mathcal{I}}_{XY}(x,y) \equiv S(P_X;\tilde{P}_X) - \langle S(P_{X|y};\tilde{P}_{X|y})\rangle_{P_Y} \tag{4.9a}$$

$$= \sum_x \sum_y p_{XY}(x,y) \log\left[\frac{\tilde{p}_{XY}(x,y)}{\tilde{p}_X(x)\tilde{p}_Y(y)}\right] \tag{4.9b}$$

This quantity can be interpreted as follows: we first model the joint probability distribution $\tilde{P}_{XY}$. We do not have elements to know if ours will be a good approximation of the real joint probability distribution yet. Suppose we also know the real probability distribution $P_X = p_X(x) = \sum_y p_{XY}(x,y)$. We can thus calculate the cross entropy $S'(P_X;\tilde{P}_X)$. Then, the information the probability distribution P_Y adds to our knowledge is the cross-mutual information $\tilde{\mathcal{I}}_{XY}(x,y)$.
The clearer manner to show what this quantity is calculating is probably by relating it to Eq. 4.3:

$$\tilde{\mathcal{I}}_{XY}(x,y) = \langle I(\tilde{P}_X)\rangle_{P_X} - \langle\langle I(\tilde{P}_{X|y})\rangle_{P_{X|y}}\rangle_{P_Y} \tag{4.10}$$

The cross-mutual information is thus the difference between the expectation value of the information we get by knowing the tentative probability distribution $\tilde{P}_X$ and the expectation value of the information we get by knowing the same probability distribution, given a certain value of y. The greater this quantity is, the more information will $\tilde{P}_{X|y}$ contain with respect to $\tilde{P}_X$. Since the two terms are definite positive, for fixed P_X and $\tilde{P}_X$ the only way to maximize the cross-mutual information is to reduce as much as possible the second term, i.e. the entropy of $\tilde{P}_{X|y}$. For this to happen, two conditions must be met:

- The knowledge of each event $y \in Y$ must select a subset of elements of X as small as possible, to minimize the entropies $S(P_{X|y})$;

- The expected and real conditional probability distributions must be as similar as possible: $P_{X|y} \approx \tilde{P}_{X|y}\ \forall y$.

The cross-mutual information is still symmetric in X and Y. Indeed,

$$\tilde{\mathcal{I}}_{XY}(x,y) \equiv S(P_X;\tilde{P}_X) - \langle S(P_{X|y};\tilde{P}_{X|y})\rangle_{P_Y} \tag{4.11a}$$

$$= S(P_X) + D_{KL}(P_X||\tilde{P}_X) - \langle S(P_{X|y}) + D_{KL}(P_{X|y}||\tilde{P}_{X|y})\rangle_{P_Y} \tag{4.11b}$$

$$= \mathcal{I}_{XY}(x,y) + D_{KL}(P_X||\tilde{P}_X) + D_{KL}(P_Y||\tilde{P}_Y) - D_{KL}(P_{XY}||\tilde{P}_{XY}) \tag{4.11c}$$

The last line clearly shows the symmetry.
Note also that, unlike the mutual information, the cross-mutual information can be negative: indeed, while it is true that $S(P_X) \geq \langle S(P_{X|y})\rangle_{P_Y}$, it is not true that $S(P_X;\tilde{P}_X) \geq \langle S(P_{X|y};\tilde{P}_{X|y})\rangle_{P_Y}$.

4.1.2 Application of the cross-mutual information to BACH

We will now apply the formalism we described to our specific case. Let $\tilde{P}_{AB,C} = \tilde{p}_{AB,C}(ab,c)$ be the distribution probability of a contact of type c between a couple of types of amino acids ab observed in the learning database TOP500, and $P_{AB,C} = p_{AB,C}(ab,c)$ be the real probability distribution observed for a certain conformation of which we want to assess the quality. Since we never observed that specific conformation before, but we know its amino acid sequence as it does not vary from a pose to another, we want to measure the cross information flux given by observing a certain contact c, knowing the probability distribution $P_{AB} = p_{AB}(ab)$:

$$I_{P;\tilde{P}}(c \to ab) = S(P_{AB};\tilde{P}_{AB}) - S(P_{AB|c};\tilde{P}_{AB|c}) \tag{4.12}$$

where $S(P_{AB};\tilde{P}_{AB})$ is the cross entropy of the probability distribution of the couples of types of amino acids ab, as it was calculated on the training set. By calculating the cross-mutual information we obtain:

$$\tilde{\mathcal{I}}_{AB,C}(ab,c) = S(P_{AB};\tilde{P}_{AB}) - \sum_c p_C(c) S(P_{AB|c};\tilde{P}_{AB|c}) \tag{4.13a}$$

$$= \sum_{ab}\sum_c p_{AB,C}(ab,c) \log\left[\frac{\tilde{p}_{AB,C}(ab,c)}{\tilde{p}_{AB}(ab)\tilde{p}_C(c)}\right] \tag{4.13b}$$

$$= -\sum_{ab}\sum_c p_{AB,C}(ab,c)\epsilon^c_{ab} \tag{4.13c}$$

$$= -\frac{1}{N}\sum_{ab}\sum_c n^c_{ab}\epsilon^c_{ab} \tag{4.13d}$$

$$= -\frac{E_{\text{pair}}}{N} \tag{4.13e}$$

where E_{pair} is the BACH pairwise score as presented in Section 2.2, ϵ^c_{ab} is the BACH parameter relative to the couple of types of amino acids ab and the type of contact c, and N is the total number of contacts of the target structure. N is constant for every conformation of a same protein, due to the presence of the class of non-contact.
Let us then rephrase in this context the explanation given in Section 4.1.1:

1. In the training phase, we construct our tentative joint probability distribution $\tilde{P}_{AB,C}(ab, c)$ and we store the information in the parameters ϵ^c_{ab}.

2. When assessing the quality of a conformation in a decoy set, we suppose to know the actual probability distribution $p_{AB}(ab)$: indeed, we know the sequence of the protein and we want to discriminate the best conformation (more appropriately, the *contact map*).

3. The score is thus proportional to how much information adds the knowledge of the contact map contained in the true probability distribution $p_C(c)$, given that we are basing our estimation on another probability distribution $\tilde{p}_{AB,C}(ab, c)$.

We point out that in this context, the more information we get, the more the test structure complies with our expectations, and thus has chances to be correct.

4.1.3 Ranking the scores

Now that we reformulated our score in terms of information flows, we are interested in understanding what makes a certain conformation "1" be scored lower than another conformation "2", given a set of conformations of a same protein. We find that, if $E^1_{\text{pair}} < E^2_{\text{pair}}$,

$$S(P^1_{AB}; \tilde{P}_{AB}) - \langle S(P^1_{AB|c}; \tilde{P}_{AB|c})\rangle_{P^1_C} > S(P^2_{AB}; \tilde{P}_{AB}) - \langle S(P^2_{AB|c}; \tilde{P}_{AB|c})\rangle_{P^2_C} \tag{4.14a}$$

$$\langle S(P^1_{AB|c}; \tilde{P}_{AB|c})\rangle_{P^1_C} < \langle S(P^2_{AB|c}; \tilde{P}_{AB|c})\rangle_{P^2_C} \tag{4.14b}$$

$$\sum_c p^1_C(c) \left[S(P^1_{AB|c}) + D_{AB|c}(P^1||\tilde{P})\right] < \sum_c p^2_C(c) \left[S(P^2_{AB|c}) + D_{AB|c}(P^2||\tilde{P})\right] \tag{4.14c}$$

where $D_{AB|c}(P^i||\tilde{P}) \equiv D_{KL}(P^i_{AB|c}||\tilde{P}_{AB|c})$ for $i = 1, 2$. The inequality 4.14a reduces to 4.14b because it is always true that $P^1_{AB} = P^2_{AB} \Rightarrow S(P^1_{AB}; \tilde{P}_{AB}) = S(P^2_{AB}; \tilde{P}_{AB})$.
We can conclude that there are two factors which contribute to associating a better score to conformation "1" rather than "2":

- $\langle S(P^2_{AB|c})\rangle_{P^2_C} > \langle S(P^1_{AB|c})\rangle_{P^1_C}$: the probability distributions of the couples of amino acids (given they are in a certain class of contact c) relative to conformation "1" convey on average more information than the probability distribution relative to conformation "2"

- $\langle D_{AB|c}(P^2||\tilde{P})\rangle_{P^2_C} > \langle D_{AB|c}(P^1||\tilde{P})\rangle_{P^1_C}$: the average of the Kulback-Leibler divergence of those same distributions from the probability distribution observed in the learning database is larger for conformation "2" than for conformation "1".

These are analogous to the two conditions we pointed out at the end of Section 4.1.1.
The two factors grasp different aspects: if on average the information gained with the knowledge of the contact class c is higher, it means that there are differences among the kinds of contacts formed by the different residues. For example, if in one test conformation the "α-helix contact" shows to be almost equiprobable among all the couples of residue types, while in another it shows to prefer a subset of them, the best score will be assigned to the latter conformation. But, what if the subset of couples of residue types selected is wrong? This term will equally prefer the more selective conformation. However, the second term will disfavor that conformation, since the distribution $P^{\text{wrong}}_{AB|c}$ will be very different from the distribution $\tilde{P}_{AB|c}$ observed in the training database.

In order to score different configurations of known sequence and unknown contact map, it is more convenient and straightforward to think in terms of conditional probabilities of the form $P_{C|ab}$ rather than of the form $P_{AB|c}$. We note that instead of Eq. 4.14a we could have written

$$S(P^1_C; \tilde{P}_C) - \langle S(P^1_{C|ab}; \tilde{P}_{C|ab})\rangle_{P^1_{AB}} > S(P^2_C; \tilde{P}_C) - \langle S(P^2_{C|ab}; \tilde{P}_{C|ab})\rangle_{P^2_{AB}} \quad (4.15a)$$

$$\langle S(P^1_{C|ab}; \tilde{P}_{C|ab}) - S(P^2_{C|ab}; \tilde{P}_{C|ab})\rangle_{P_{AB}} < S(P^2_C; \tilde{P}_C) - S(P^1_C; \tilde{P}_C) \quad (4.15b)$$

where we put $P_{AB} \equiv P^1_{AB} = P^2_{AB}$. This equation is dependent on the unknown cross entropies of the distributions of contacts of the two conformations consid-

ered, unless $P_C^1 \sim P_C^2 \Rightarrow S(P_C^1; \tilde{P}_C) \sim S(P_C^2; \tilde{P}_C)$. This condition is met both in protein folding and in protein-protein interaction for conformations that do not differ significantly: for example, for all the conformations in the vicinities of the native state, and in protein-protein interaction for basically all possible poses, since the only contacts that vary are the few at the interface. For these classes of configurations, if $E^1_{\text{pair}} < E^2_{\text{pair}}$ at least one of these two conditions must be met:

- $\langle S(P^2_{C|ab}(c|ab)\rangle_{P_{AB}} > \langle S(P^1_{C|ab}(c|ab))\rangle_{P_{AB}}$: the probability distributions of the contacts relative to conformation "1" convey on average more information than the same probability distributions relative to conformation "2"

- $\langle D_{C|ab}(P^2||\tilde{P})\rangle_{P_{AB}} > \langle D_{C|ab}(P^1||\tilde{P})\rangle_{P_{AB}}$: the average of the Kulback-Leibler divergence of those distributions from the probability distribution observed in the learning database is greater for conformation "2" than for conformation "1".

The first of these two conditions tells something qualitatively different from the previous ones: to be scored better than another one, a conformation must display, on average, a smaller entropy in the distribution of contacts once the couple of types of amino acid is known. Thus we are saying that, if $P_C^1 \sim P_C^2$ holds, the reverse of what we stated when commenting Eq. 4.14b is also valid: on average, the more selective is the given couple of amino acids on the set of contact classes, the better score the conformation will get. It follows that the scoring function favors conformations displaying specific kinds of interactions, assuming that these interactions are among the set of contact classes considered (see property 3. of the mutual information estimator, reported in Section 4.1.1, which still holds for the cross-mutual information). We will see the implications of this inequality in the next section.

4.2 Refining statistics

4.2.1 Upper and lower bounds of the cross-mutual information $\tilde{\mathcal{I}}_{AB,C}(ab, c)$

Whatever probability distribution we may choose to bias our estimator, it always holds that $\tilde{\mathcal{I}}_{AB,C}(ab, c) \leq S(P_C; \tilde{P}_C)$. This follows directly from applying the

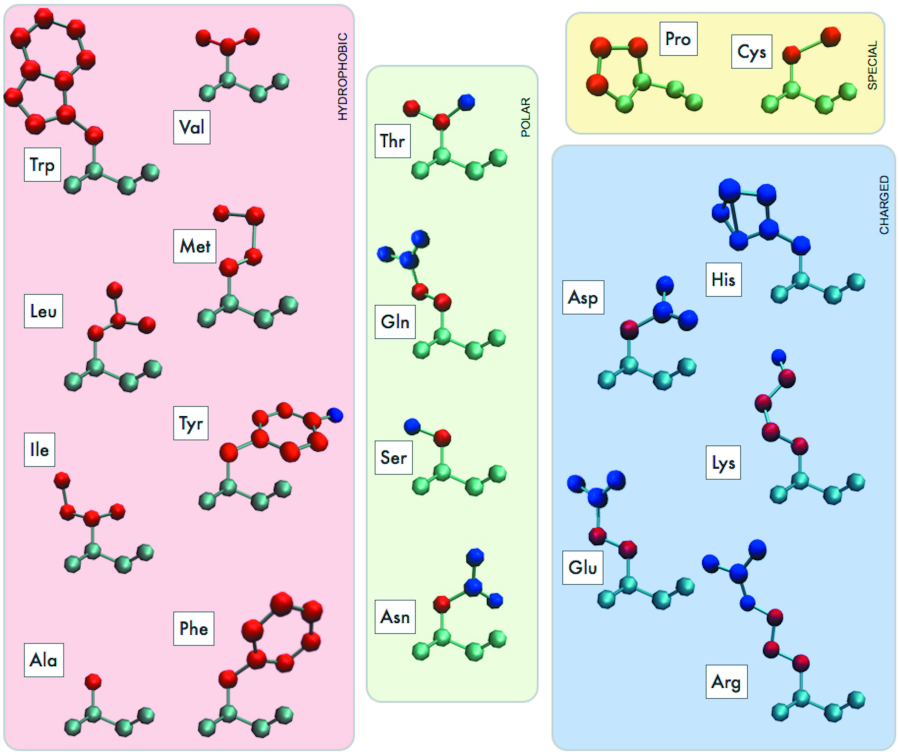

Figure 4.1: The twenty amino acids and the polarity associated to each atoms of their side chains: red=apolar; blue=polar.

$ab \leftrightarrow c$ symmetry property to the definition of cross-mutual information given in Eq. 4.13a, because the second term is always positive. The inequality implies that the information we can have by acquiring data may never be greater than the entropy of the distribution of the data. For the cross-mutual information an even stricter bound holds, since the term $\sum_{ab} p_{AB}(ab) S(P_{C|ab}; \tilde{P}_{C|ab})$ cannot be zero unless $P_{C|ab} = \tilde{P}_{C|ab}\ \forall ab$ and the entropy of each probability distribution $\tilde{P}_{C|ab}$ is null, which is impossible, as it would require to have as many elements in the set of contact types C as in the set of couple of residue types AB. The second term is minimized only if $P_{C|ab} = \tilde{P}_{C|ab}\ \forall ab$: thus the generalized mutual information is upper-bounded by

$$\tilde{\mathcal{I}}_{AB,C}(ab, c) \leq S(P_C; \tilde{P}_C) - \sum_{ab} p_{AB}(ab) S(P_{C|ab}) \tag{4.16}$$

To define the lower bound of the estimator, we recall that, unlike the ordinary mutual information, it can assume negative values. This means that the knowledge of a second probability distribution may cause a loss of information, if the observed data are in net contrast with the ones acquired from the learning

database. The lower bound can be defined quite simply by looking at Eq. 4.13c, as the case in which all contacts observed are of the least favorable type:

$$\tilde{\mathcal{I}}_{AB,C}(ab,c) \geq -\max_{ab,c} \epsilon^{c}_{ab} = -\max_{ab,c} \log\left[\frac{\tilde{p}_{AB,C}(ab,c)}{\tilde{p}_{AB}(ab)\tilde{p}_{C}(c)}\right] \tag{4.17}$$

4.2.2 Extending the upper bound

Adding a new contact class in BACH is likely to increase the upper bound of the estimator. Indeed, this would increase both terms of Eq. 4.13a. Yet, if the additional class is appropriately chosen, the majority of residues will prefer one of the two classes over the other. Thus we expect that the term $\sum_{ab} p_{AB}(ab) S(P_{C|ab}; \tilde{P}_{C|ab})$ will increase more than the term $S(P_C; \tilde{P}_C)$.
We also remark that this additional "segregation" goes in the direction of the requirements we stated in Section 4.1.2.

4.2.3 Reducing the noise of the estimator: polar/apolar contact classes

The most important contribution to the accuracy of the scoring function is that of reducing the noise, namely the false-positive cases of favorable and unfavorable interactions. We here present an heuristic motivation based on physical and chemical observations on the quality of the contacts defined by BACH.
The simple classification used in [CGL+12, SZC+13] is effective for distinguish-

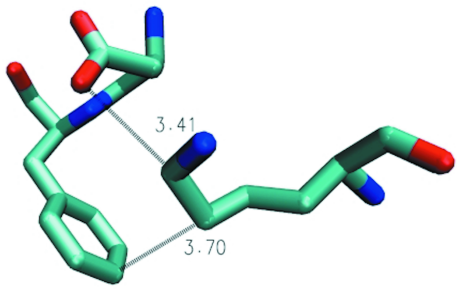

Figure 4.2: A case in which a hydrophilic residue (LYS) and a hydrophobic one (PHE) make a hydrophobic contact at the interface between two subunits, here colored in blue and red. To this contact a favorable energy value should be associated.

ing the folded state of a globular protein among a set of misfolded structures,

but is not sufficiently detailed to capture more subtle structural differences, like those observed between putative complex structures of the same dimer. An important detail that is not accounted for in the original functional form is that some residues are large and inhomogeneous with respect to their chemical properties. An example is shown in Fig. 4.2, where the same residue, LYS, forms a hydrophobic contact with the hydrophobic residue PHE, and a hydrogen bond with the hydrophilic residue ARG.
Clearly, this configuration is favorable, since no hydrophobic group is in contact with a hydrophilic group. However, in the original formulation of BACH this configuration is not distinguished by a case in which, for example, PHE competes with ARG for a contact with the polar part of LYS, which of course is a less stable conformation. Since cases in which polar residues can make both polar and hydrophobic contacts are far from being uncommon, we split the sidechain-sidechain contact class into two distinct ones, which account for apolar and polar contacts. The algorithm first checks the polarity of the two atoms which make the closest contact between the considered residue pair. If at least one of them is polar, then the contact is defined polar. Otherwise, if both are apolar (namely are alyphatic carbons or sulfur atoms), the contact is defined as apolar. Two residues make contact if the distance between their two closest sidechain heavy atoms is less than 4.5, as in the original version. In BACH-S, we consider polar and apolar contacts only for residue pairs separated at least five along the sequence. The splitting between polar and apolar contacts does not affect all the parameters, since $p_C(c_{old}^{VdW}) = p_C(c_{new}^{apolar}) + p_C(c_{new}^{polar})$. For the sake of clarity, in Fig. 4.1 is reported the complete list of residue atoms and their polarity.

As we see, there are many couples of of amino acids which can only perform one of the two types of contact. For them, the advantage described above does not count and the noise in the score remains unchanged. Although it might seem a waste, it is not so: in the previous section we stated that whenever $P_C \sim \tilde{P}_C$ can apply, a conformation "1" would be scored more favorably of a conformation "2" if $\langle S(P^2_{C|ab})\rangle_{P_{AB}} > \langle S(P^1_{C|ab})\rangle_{P_{AB}}$. This is true in the case that the average frequencies of two contact classes differ more: for the two new classes of polar and apolar VdW contact, this is true when there is a net preference for one kind of those contacts. This could help in recognizing hydrophilic hotspots, close packings of apolar atoms near the interface

or hotspots stabilized by charged and polar residues. In protein folding, the same effect could help inducing the formation of a hydrophilic bulk of the protein. Although this condition tends to favor radically polar or apolar contact maps, it does not favor unphysical conformations over physical ones: its counterpart $\langle D_{C|ab}(P^2||\tilde{P})\rangle_{P_{AB}} > \langle D_{C|ab}(P^1||\tilde{P})\rangle_{P_{AB}}$ balances it by preventing the algorithm from choosing contact distributions too different from the one observed in the learning database.

4.3 Accounting for clashes

In the original version of BACH, it is implicitly assumed that all the structures that are compared are viable protein structures, without sizable steric clashes between atoms. This condition is verified by most of the decoys presented for the CASP competition, which are normally refined by molecular dynamics or similar approaches prior to submission. Conversely, in protein-protein interaction studies many docking algorithms typically produce structures compenetrating each other, as a result of the optimization of scores based on interface contacts. The energy terms of the original BACH algorithm are not able to deal with this issue: the pairwise contact term will associate to the compenetrated structures a more favorable energy, because of the larger number of contacts. Therefore, we introduced a term in the energy function that penalizes the clashes among residues. This penalty term is a quadratic function of the distance between two atoms, when this is smaller than a certain threshold, and identically zero otherwise. The threshold and the force constant of the quadratic penalty depend on the species of the two atoms, but not on residue type. We derived these parameters by computing, as a function of the distance, the distributions of contact frequencies between atom pairs of given elements on TOP500, the same data set of single chain globular proteins used for deriving the other parameters. The logarithm of these frequency distributions up to its first maximum is then fitted to a quadratic curve, as can be seen in Fig 4.3. We also verified that very similar frequency distributions are found when we use a dataset of homo- and heterodimers [HVJW10]. Thus, derived parameters do not depend on whether single chain globular proteins or protein complexes are used as datasets.

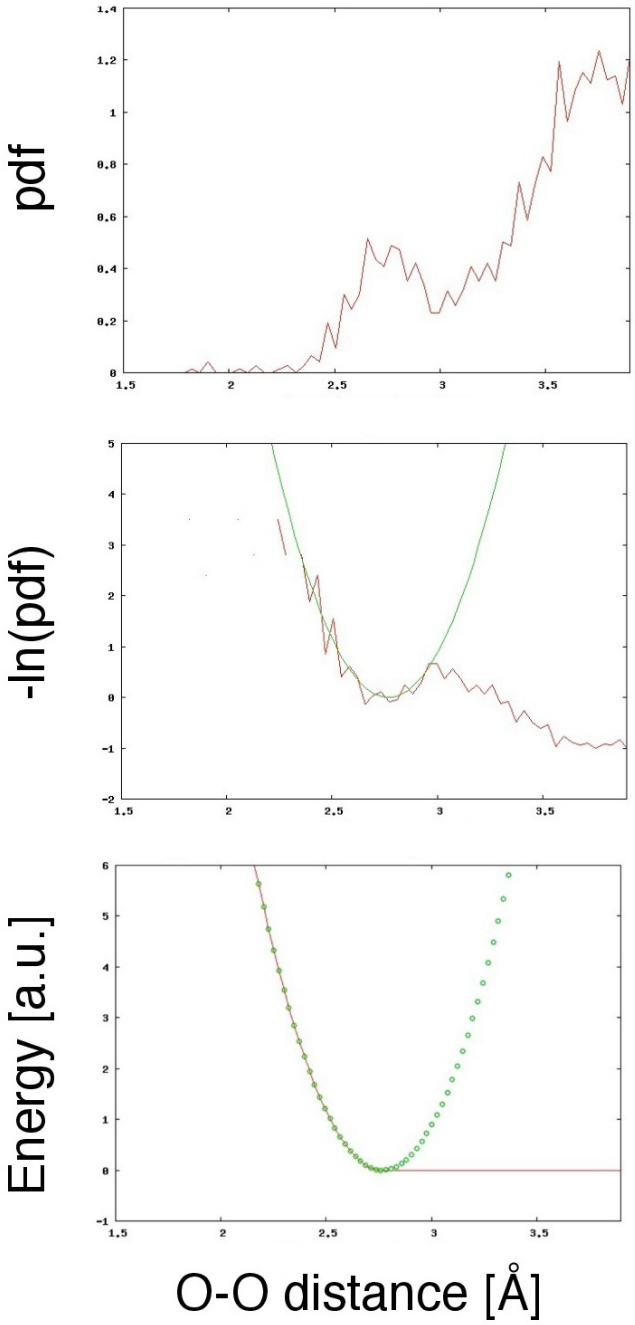

Figure 4.3: a) To calibrate the parameters of the quadratic clash term, the frequency distribution of distances between two fixed types of atoms (here oxygen and oxygen) is computed. b) Then, the logarithm is taken and the tail for small distances of the distribution is fitted with the quadratic curve. c) The clash term will then be zero for distances greater than a threshold determined by the x-coordinate of vertex of the fitted parable, and will grow quadratically for smaller distances, according to the fit.

The new BACH energy function can then be written as

$$E_{\mathrm{BACH}} = p_1 E_{\mathrm{PAIR}} + E_{\mathrm{SOLV}} + p_2 E_{\mathrm{CLASH}}. \tag{4.18}$$

We keep $p_1 = 0.6$, used for the pairwise term in [CGL$^+$12, SZC$^+$13]. The new clash penalty term is weighted by a prefactor in order to have the same standard deviation of the pairwise term on the decoy sets analyzed. To avoid overfitting, we calibrated this parameter on a decoy set from CASP 9 that was not included in the test set (T0629), this gave $p_2/p_1 = 0.03$. Subsequently, we checked that this choice was consistent also with other decoy sets used in our tests.

4.4 Testing on CAPRI and CASP decoy sets

The need of reliable test sets to assess the quality of potentials for protein-protein interaction has driven the community to set up several databases of decoy sets and several docking algorithms able to produce challenging conformations. As we will see in Chapter 5, the choice of relying on only one algorithm for the production of the decoy structures always creates a bias on the scoring functions quality assessment. In order to avoid this, to validate our scoring function we look for decoy sets created in the most inhomogeneous way. CAPRI [JHM$^+$03] is a popular community-wide experiment which counts more than 30 separate competition rounds and more than 60 targets. We obtained by the competition organizers 26 targets, and, among these, we selected the 17 characterized by a number of chains equal to two and the absence of RNA, DNA or other non-amino-acid-based molecules. We merged the couples of decoy sets corresponding to Target 11 and 12, Target 24 and 25 and Target 35 and 36, since for each of these couple of targets the same complex structure was to be predicted. The complete list of target decoy sets used in this thesis, along with the number of poses contained in each one, is provided in Tab. 4.1. Importantly, we also checked that the performance of BACH was maintained also for protein folding problems. We thus tested the new version of the algorithm against the 33 CASP 8-9 decoy sets used in the tests for the original version of the method. For protein folding problems, we applied the same procedure described in [CGL$^+$12] and in Chapter 2.

Table 4.1: The CAPRI targets together with the corresponding PDB structure code, the number of docking poses in the decoy set (used for the native pose recognition test), the native rank for the different scoring functions analyzed in this thesis, and the number of poses selected for the model quality tests (fraction enrichment and lowest iRMSD within best 30 poses).

Target	PDB code	Total decoys	BACH-S	PIE*PISA	IRAD	Rosetta	HADDOCK	FireDock	Decoys for quality assessment
(11+12)	1ohz	405	1	12	43	33	55	73	405
22	1syx	77	2	4	6	15	4	17	77
23	2bbw	84	46	1	8	45	37	5	84
(24+25)	2j59	751	26	48	213	350	328	51	532
26	2hqs	427	54	16	87	112	96	60	405
28	2oni	453	28	93	96	86	162	158	399
29	2vdu	526	118	54	202	280	179	56	495
32	3bx1	474	298	103	155	331	134	82	474
(35+36)	2w5f	839	7	30	256	19	146	165	681
41	2wpt	457	17	127	112	80	52	117	457
46	3q87	557	6	29	114	76	141	27	537
47	3u43	405	71	24	129	20	120	39	370
53	4jw2	514	17	58	61	79	91	11	507
54	4jw3	507	89	150	211	104	224	311	496

4.4.1 Interface BACH-SixthSense score

We also calculated scores by only summing the contributions of the residues at the interface between the two subunits. This measure is interesting for two reasons: first, because many of the analyzed methods only take into consideration interfaces, second because by restricting to that specific region the approximation $P_1(c) \sim P_2(c)$ behind the inequality Eq. 4.15b does not hold anymore, and provide us with an estimation of the difference between relying only on the two conditions given by Eq. 4.14c and being able to rely also on the other two conditions expressed by Eq. 4.15b. Lacking these, we expect a decrease in the performance. Interface scores are computed using the pairwise and clash terms for all residue pairs, that involve interface residues. Interface residues are those with a heavy atom within 10 Å of distance from any heavy atom of the other subunit.

4.4.2 Comparison with other scoring functions

We compared the performances of BACH-S with five scoring functions that achieved good results in protein-protein interaction in the last few years: IRAD, PIE*PISA, Rosetta, HADDOCK and FireDock. In Chapter 2 we provided a detailed description of these scoring functions. IRAD [VHW11] is the latest version of the well-known ranking algorithm ZRANK [PW07]. The original scoring function was introduced to rank the structures predicted by the docking

program ZDOCK [PHW11]. It includes Van der Waals and electrostatic terms calculated using parameters from the CHARMM19 [BBO+83] force field, together with statistical contact potentials, atomic and residue based, the latter of whom is taken from I-TASSER algorithm [Zha08]. PIE [DE10] and PISA [VRE13] are more recent scoring functions, founded on a residue-based and an all-atom description, respectively. The Van der Waals potential is modeled, in PISA, through the OPLS [JTR88] force field, while the contact potential, in PIE, is defined through a sum over residue pairs of a smoothed step function of their distance with a set of residue-dependent parameters, learned on a training set of native structures of protein complexes together with different associated sets of decoy structures. The Rosetta scoring function [GMW+03] is used in one of the most popular methods for native state recognition and protein docking. The scoring function is built by exploiting Bayes' theorem to determine the probability of having a certain structure of a protein complex, given the amino acid sequence and the structure of the monomers. It also accounts for the interaction with the solvent by using an implicit solvation score according to the model of Lazaridis and Karplus [LK99]. We described in detail the Rosetta statistical foundations in Chapter 2. HADDOCK [DBB03] and FireDock [ANW07] are scoring functions extracted from the corresponding docking programs. They are not available as standalone programs, but the authors use it in CAPRI scoring competitions, obtaining excellent results [JHM+03]. HADDOCK energy function is a weighted sum of physical terms that keep into account Van der Waals and electrostatic forces, geometrical restraints, solvation and binding energies. Similarly, FireDock energy function implements almost the same variety of interactions. For the analysis of the CASP decoy sets for monomeric proteins we have used the same scoring functions analyzed in [CGL+12]: QMEAN6 [BTS08, BST09, BBS11], RF_CB_SRS_OD [RF10] and Rosetta [TBM+03], which we had already considered as representative of the state-of-the-art for single chain native state discrimination. It is important to note that, within the Rosetta framework, the parametrizations used for evaluating alternative docking models or the single chain CASP targets are different.

4.4.3 Assessing the performance in CAPRI decoy sets: native pose discrimination

An important technical difficulty in working with CAPRI decoy sets is related to the great extent of structure inhomogeneity present for the same target. The number of residues in different complex structures, within the same decoy set, is often uneven, while scoring functions based on any kind of extensive quantity need to compare structures with an equal, or at least similar, number of fundamental units (these being atoms or residues depending on the scoring function). A rough energy score ranking would thus favor the structures having more residues, neglecting the conformation-dependent features of the prediction. In the CASP decoy sets used in [CGL$^+$12, SZC$^+$13] and in this thesis, targets were selected and decoy sets filtered in order to fairly compare the scores only of structures sharing the same residues. A similar filtering would essentially strip the CAPRI decoy sets. Thus, we propose a different way to overcome this difficulty. In order to compare the native pose with each decoy in the set, only the residues present in both the native and the considered decoy pose are selected. We checked that this procedure never results in cutting more than 20% of one of the structures. Then, we compute the score only for the selected residues. This allows deciding unambiguously if a docking pose has a higher or lower score than the native one, even when there are missing residues in one or both of them. The native normalized rank in a given decoy set is then given by one plus the number of times that the native pose scores less favorably than a decoy, divided by the total number of structures in the set, including the native. As usual, the lower the rank, the better the performance. For a set of homogeneous structures, the rank described above is just the usual one. The same strategy is used for all the scoring functions analyzed in this thesis.

4.4.4 Assessing the performance in CAPRI decoy sets: fraction enrichment and best pose selection

In order to evaluate the quality of the decoy poses as selected by BACH-S best scores, we need to use estimators that can be computed without any knowledge of the native structure. To this aim, we first analyzed the fraction enrichment, calculated as the number of good structures as a function of the fraction of best-scored structures considered. The quality of the structures is assessed via the standard CAPRI criteria, by evaluating iRMSD (interface RMSD), lRMSD

(ligand RMSD), and fnat (fraction of native interface contacts) [MLLW05]. In order to assess the quality of the best poses selected by a scoring function, we then computed the minimum iRMSD among the best 30 scored structures. This number corresponds to approximately one tenth of the average number of decoys. Within the perspective of exploiting a scoring function to select a set of models for further refinement, this is roughly the number of structures that one can viably consider.
As already discussed in the previous section, the performance of scoring functions based on extensive quantities defined on the whole complex structure, such as BACH-S, may be biased by the presence of significant differences in residue numbers within the compared poses. Indeed, the above estimators (fraction enrichment and lowest iRMSD within the best 30 poses), as any other based on interface structural features, are not sensitive to residues away from the interface. On the other hand, scoring functions that evaluate extensively the whole structure are affected. In order to reduce this bias, we first operated a selection to avoid comparing poses with too different residue numbers. We selected the poses whose lengths lie within the most populated 25-residue long interval of lengths. Then, in computing fraction enrichment and lowest iRMSD, we considered scores rescaled by the number of residues for BACH-S and Rosetta, because they are based on the whole structure. We did not rescale the other four scoring functions because they produce interface scores. Rescaled scores are used to compute their correlations with iRMSD as well.
We checked that the considered estimators (fraction enrichment and lowest iRMSD within the best 30 poses) are robust for small changes in the selection procedure described above. The discarded decoys usually represent at most 10% of the entire decoy set (see Tab. 4.1). In just two cases we discard a higher fraction: in Target 35+36 30% and in Target 24+25 20% of the decoy set.

4.5 Results

4.5.1 Native state recognition for monomeric proteins: CASP decoy sets

We first tested BACH-S on the 33 CASP 8-9 targets [CKT09] with the same filtering of decoy sets used to test the original algorithm for monomeric proteins [CGL$^+$12, SZC$^+$13]. The results are presented in form of normalized ranks:

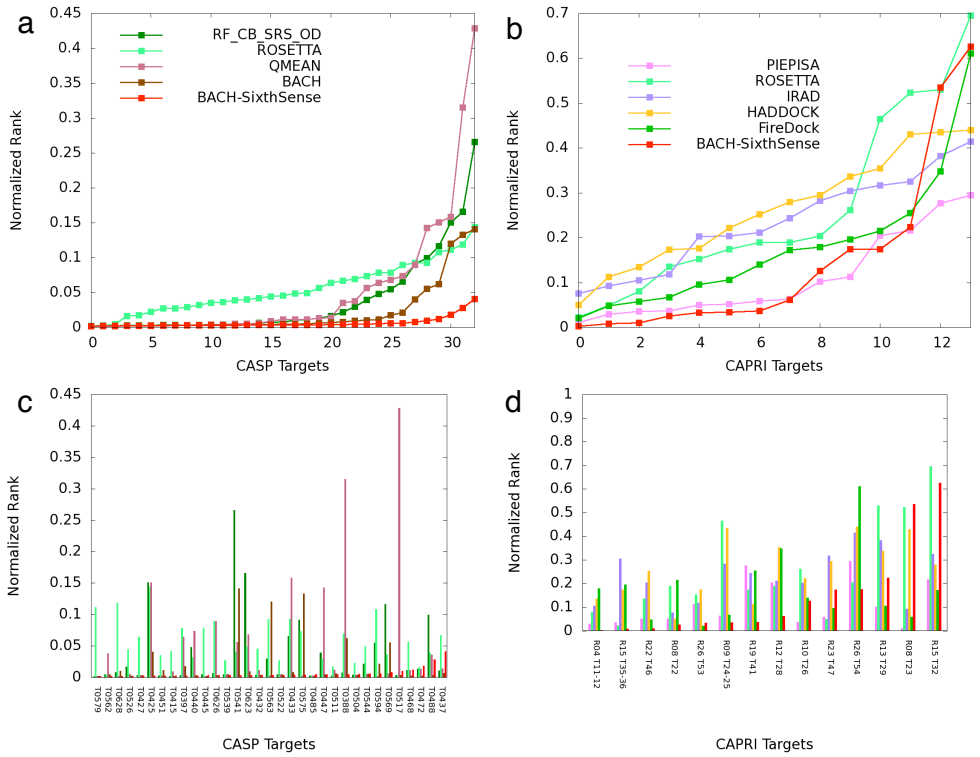

Figure 4.4: a) Normalized ranks of scoring functions for native structure prediction of monomeric proteins in CASP 8-9 decoy sets. Note that the same decoy set does not generally occupy the same position along the x-axis, since the ranks are ordered from the best to the worst independently for each scoring function. The lines are only guides for the eyes. b) Normalized ranks of scoring functions for protein-protein structure prediction in the CAPRI decoy sets listed in Tab. 4.2, in the same representation of Panel a. c), d) Same data as a), b) presented target by target. Here the decoy sets are sorted according to the performance of BACH-S

after having associated an energy score to each decoy conformation and to the native one, the scores are sorted from the lowest to the highest, and the position of the native in the rank is assessed and normalized with the number of structures in the set. The lower the rank, the better the performance. In Fig. 4.4a we report the ranks of BACH-S, along with the original version of the algorithm and the other tested competitor scoring functions for protein folding problems. We see that results from the modified algorithm even outperform the quality of the original version. In particular the native state has the lowest score and is thus correctly discriminated in 22 out of 33 decoy sets (67%).

Table 4.2: The CAPRI targets together with the corresponding PDB structure code, the number of docking poses in the decoy set (used for the native pose recognition test), the number of poses selected for the model quality tests (fraction enrichment and lowest iRMSD within best 30 poses), and the rank of the native structure for BACH-S scoring function and for only its interface contribution.

Target	PDB code	Total decoys	BACH-S	BACH-S int.
(11+12)	1ohz	405	1	21
22	1syx	77	2	13
23	2bbw	84	46	1
(24+25)	2j59	751	26	39
26	2hqs	427	54	50
28	2oni	453	28	312
29	2vdu	526	118	31
32	3bx1	474	298	149
(35+36)	2w5f	839	7	162
41	2wpt	457	17	115
46	3q87	557	6	114
47	3u43	405	71	80
53	4jw2	514	17	7
54	4jw3	507	89	191

4.5.2 Native docking pose recognition: CAPRI decoy sets

Secondly, we tested BACH-S on decoy sets of docked protein-protein complexes for native pose recognition. In order to avoid possible biases due to uneven residue numbers in the compared poses, we devised a ranking method where each pose is separately compared to the native pose, using only the residues shared by the two poses (see Section 4.4.3 for details).

In Fig. 4.4b we plot the normalized ranks for the six analyzed scoring functions on 14 CAPRI decoy sets, 3 of which consisting of two targets merged together. On these test cases, BACH-S performs significantly better than IRAD, Rosetta, HADDOCK and FireDock and marginally better than PIE*PISA. In the top panels targets are sorted in order of performance separately for each scoring function. We choose this ordering to underline the general trend of the performance of the scoring functions, which would otherwise be unintelligible. Lines are only guides for the eyes. The same data of panels 4.4a and 4.4b are presented also in panels 4.4c and 4.4d, respectively. In these panels, the ordering of the decoy sets along the x-axis is fixed, to better compare the performance of each scoring function on the same target. Details of the performances for each target are reported in Tab. 4.2. BACH-S has the best performance in 8

decoy sets out of 14 with respect to the other scoring functions. The native complex is ranked by BACH-S as first in one case, in the top 10 structures in 4 cases (29%), and is within the best 10% in 8 cases (57%). PIE*PISA has a similar performance, but unlike BACH is specific for protein-protein interaction and cannot be used also for monomeric proteins. Fig. 4.4d and Tab 4.1 show a mild correlation between the scoring functions we analyzed. This complementarity could be exploited to build a consensus scoring function, based on the combination of many. This strategy has already been explored by several groups, and the optimal combination is often far from trivial and requires an ad hoc statistical analysis (see for example QMEANclust [BST09]).

We also checked if the capability of BACH-S of discriminating the correct structure derives dominantly from a single term of the scoring function or from their combination. As shown in Fig. 4.7, the clash component and the pairwise component are able to score properly the correct structure even when used on their own, even if their combination is marginally better. The solvation component achieves instead marginally worse scores, comparable to those of IRAD and FireDock. We also verified that the performances of BACH-S are not correlated with the size of the interface or with the volume of the protein complex. We found only a mild correlation with the ratio interface/volume. Remarkably, the performance of the BACH-S scoring function is crucially based on evaluating the whole complex structure. Indeed, if only interface residue pairs are scored using the pairwise and clash terms (see Section 4.4.1 for details), the performance in native pose recognition is worse (see dedicated table Tab. 4.2).

4.5.3 Model quality assessment: CAPRI decoy sets

Finally, we tested the ability of BACH-S in recognizing "good" near-native poses within the CAPRI decoy sets. The tests described in this section were performed after filtering decoy sets, in order to compare structures with not too different residue numbers, and using scores rescaled by the number of residues for BACH-S and Rosetta (see Section 4.4.4 for details). As a result, residue numbers differ at most by 24. The number of selected decoys for each CAPRI decoy set is reported in Tab. 4.2. Remarkably, neither BACH-S nor any other scoring function we analyzed show a significant correlation between the interface-RMSD (or equivalently ligand-RMSD) of the decoys with the native complex and the corresponding scores, as we can see in Fig. 4.8 and Fig. 4.9.

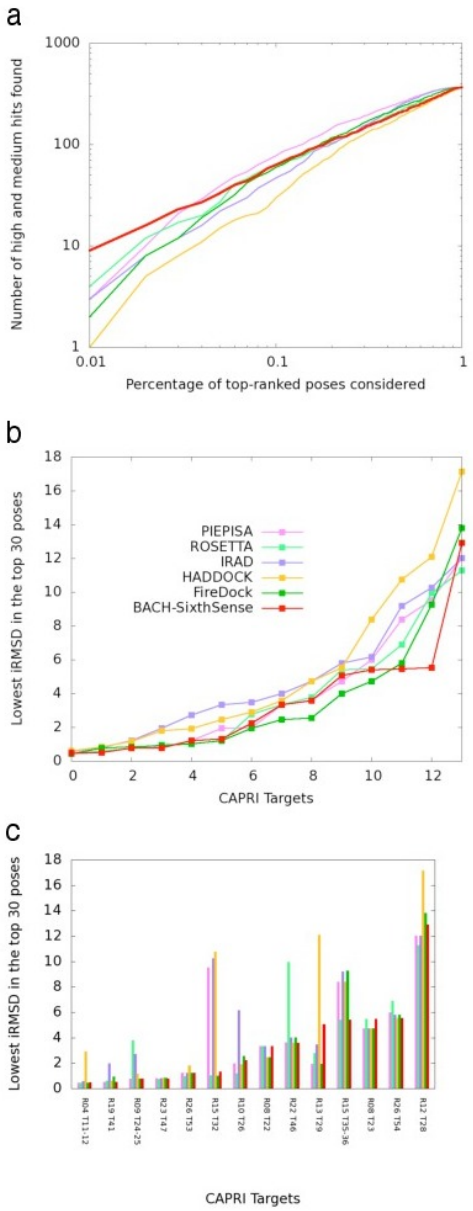

Figure 4.5: Fraction enrichment for the six scoring functions analyzed in this thesis on the 14 decoy sets reported in Tab. 4.2 for "high" and "medium" hits (a graph containing also "acceptable" hits is reported in Fig. 4.6). The quality of the structures is assessed via the standard CAPRI requirements. b) Minimum iRMSD among the 30 structures of lowest rank for each decoy set. Each line corresponds to a different scoring function. Decoy sets are sorted independently for each scoring function. Lines are only guides for the eyes. c) Same data shown in b), target by target. Here the decoy sets are sorted according to the performance of BACH-S.

Yet, significant features regarding the scoring of near-native structures can be captured by the fraction enrichment (Fig. 4.5a).

Computed as a function of the fraction of poses that are best-ranked by

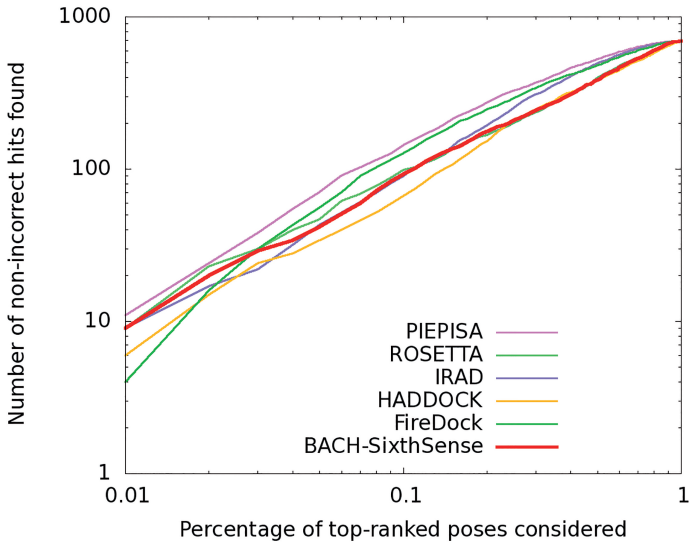

Figure 4.6: Fraction enrichment for the four analyzed scoring function on the 14 decoy sets reported in Tab. 1 for "high", "medium" and "acceptable" hits

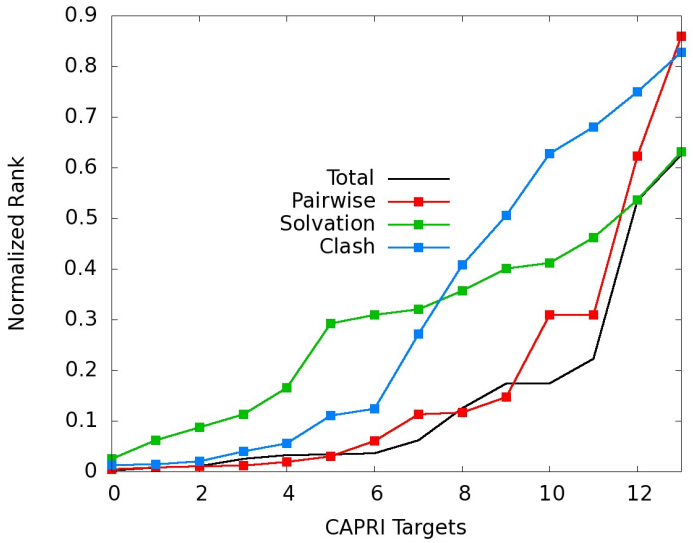

Figure 4.7: Normalized rank of the three energy contributions present in BACH-SixthSense, taken as separate scoring functions.

each scoring function, the fraction enrichment measures the number of "good" structures found, where the quality of the structures is assessed by means of the standard CAPRI criteria on iRMSD, lRMSD and fraction of native contacts fnat [MLLW05]. For structures of medium-high quality, according to CAPRI criteria, our scoring function exhibits a performance that is significantly better than all the competitors when a fraction of up to the 5% of the best-ranked poses is selected. For higher fractions of best-ranked selected poses, and, in

general, for all "good" structures (i.e. including high, medium, and low quality hits), PIE*PISA performs better.

We finally assessed the ability in recognizing near-native decoys by finding how much close-to-native a pose can be found among the top N poses as ranked by the scoring function in use. In Fig. 4.5b and 4.5c we report, for each scoring function and each decoy set, the minimum iRMSD value among the top 30 scored structures. The data are presented like in Fig. 4.4. Thus, in panel 4.5b targets are ordered differently for each curve. In this task, the best performance is achieved by FireDock, followed very closely by our scoring function.

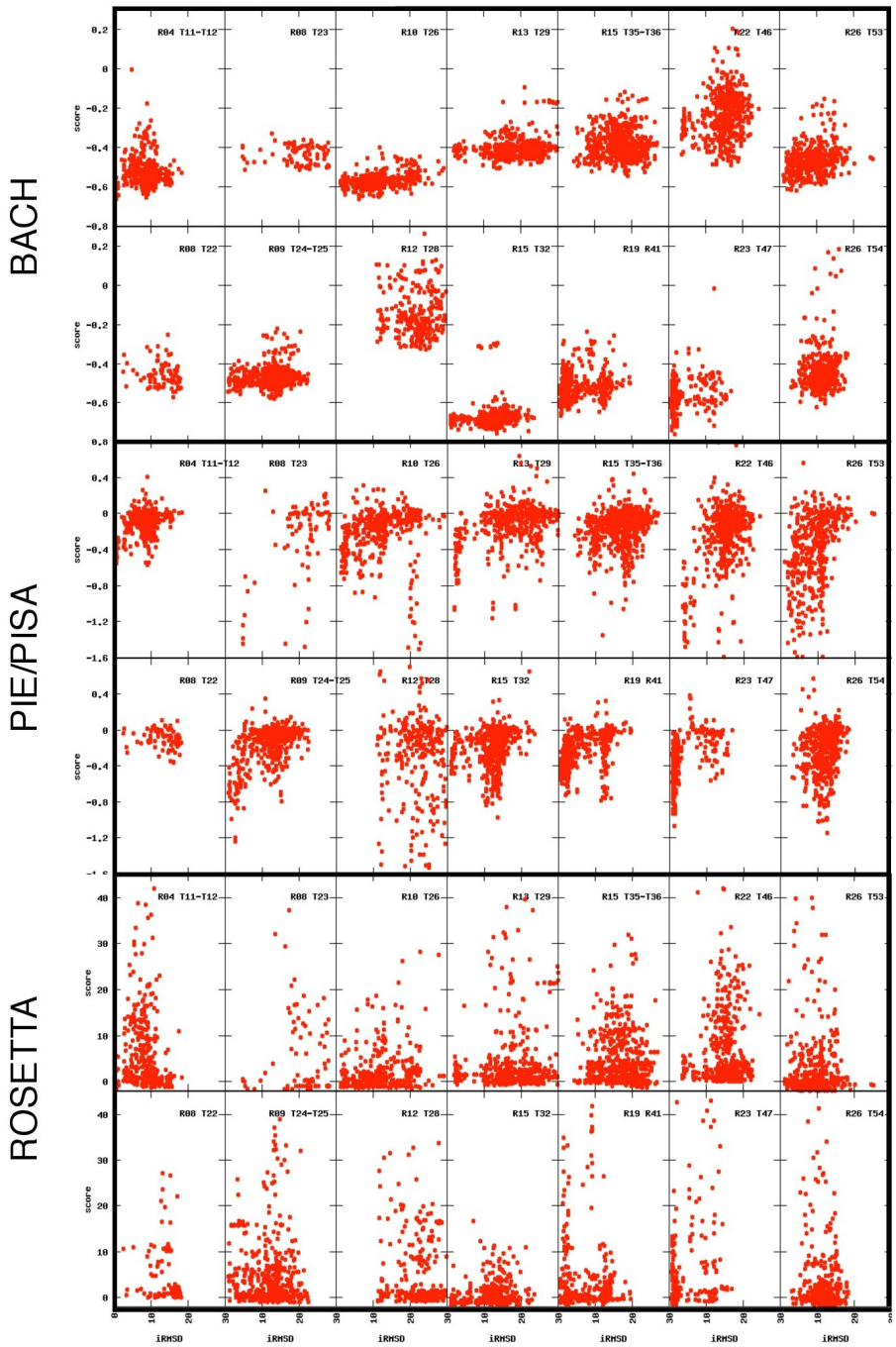

Figure 4.8: BACH, PIE/PISA and Rosetta energies as a function of the iRMSD (in Å). BACH and Rosetta energies are divided by the number of residues in each structure. Each panel corresponds to one of the 14 CAPRI decoy sets analyzed. Targets 11 and 12, 24 and 25 and 35 and 36 are merged, as explained in Section 4.4. The targets are presented in the same order of Tab. 4.2. Apart from isolate cases, no significant correlation between these two quantities is observed for any of the scoring functions.

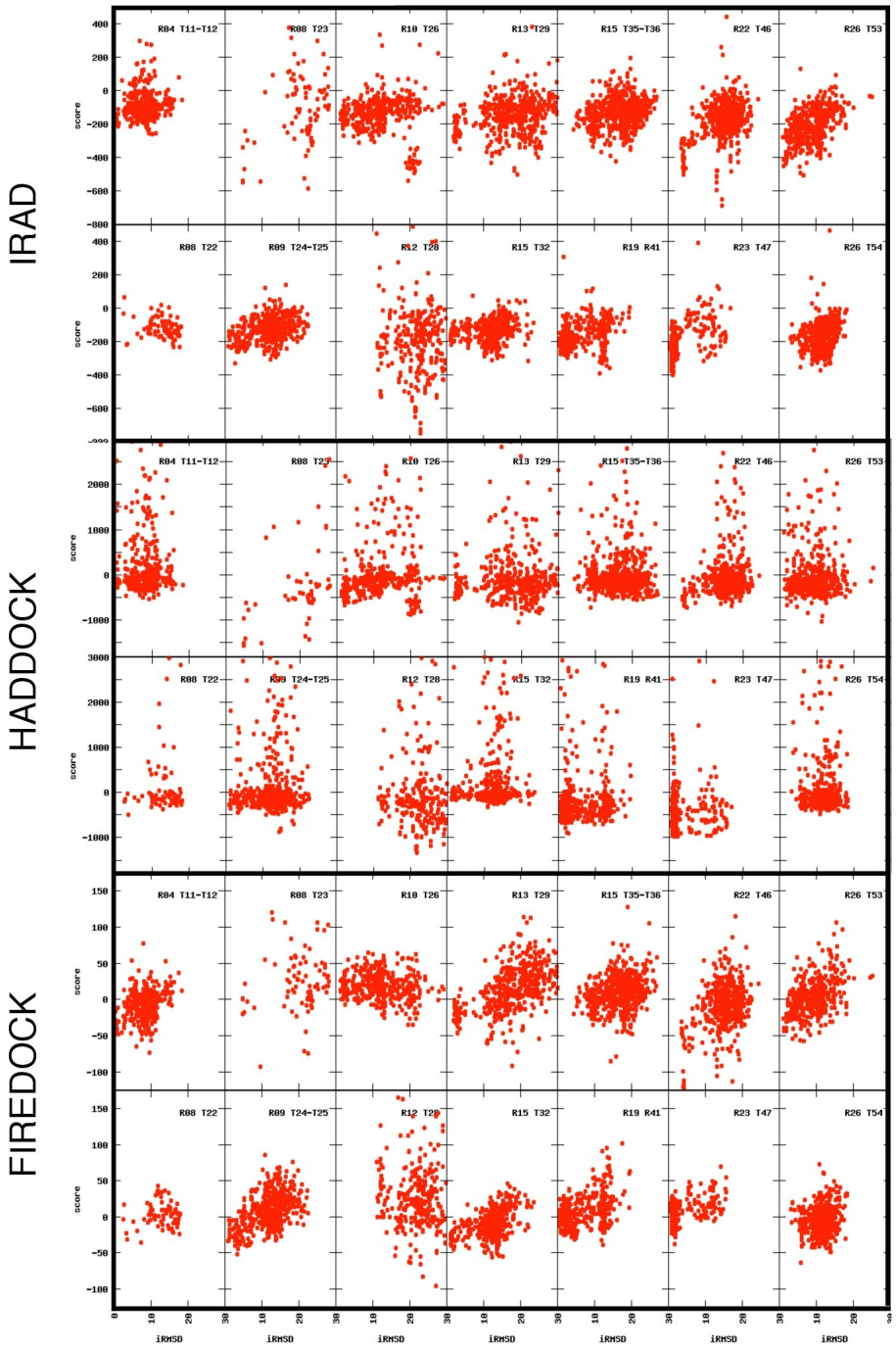

Figure 4.9: IRAD, HADDOCK and FIREDOCK energies as a function of the iRMSD (in Å). Each panel corresponds to one of the 14 CAPRI decoy sets analyzed. Targets 11 and 12, 24 and 25 and 35 and 36 are merged, as explained in Section 4.4. The targets are presented in the same order of Tab. 4.2. Apart from isolate cases, no significant correlation between these two quantities is observed for any of the scoring functions.

4.6 Discussion

BACH-S is a new version of the BACH scoring function [CGL+12, SZC+13] that can be used to evaluate protein conformations in the context of both single protein chains and protein-protein complexes. Two distinctive features of BACH-S are the splitting of polar and apolar sidechain contacts into two different classes, and a potential energy term disfavoring steric clashes. Including these modifications is crucial in the protein-protein interaction problem, and significantly improves the performance also for monomeric proteins.

4.6.1 Information flow

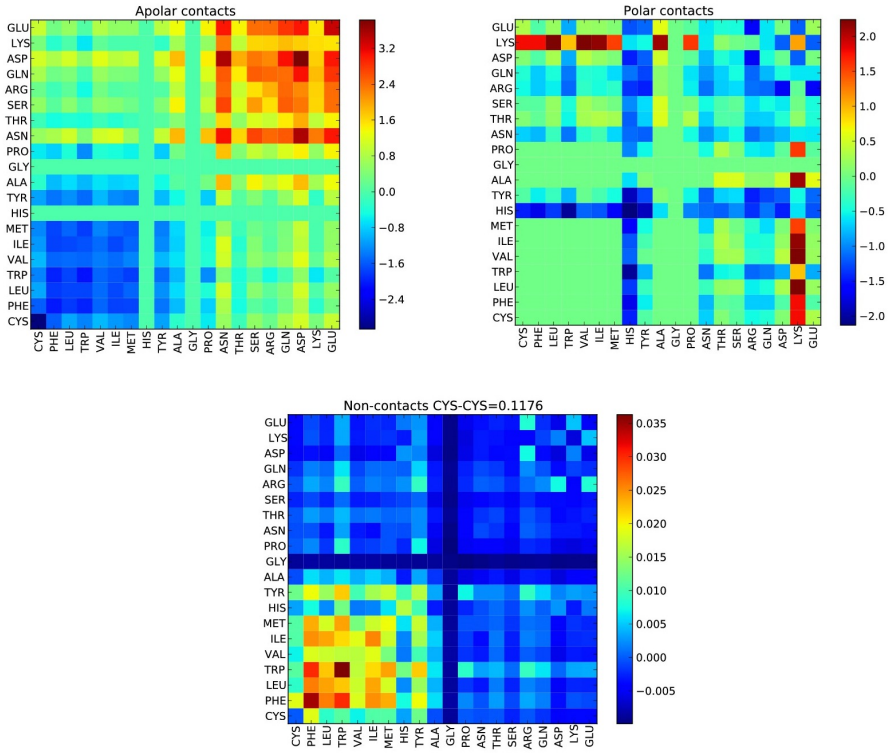

Figure 4.10: Magnitude of BACH-S parameters for the polar, apolar and non-contact classes. The residues are sorted by decreasing hydrophobicity, from left to right and from bottom to top.

From the theoretical point of view we provided a novel framework to rationalize the statistical method used in BACH. The gain in performance attained by splitting the Van der Waals contact class into polar and apolar contact classes is supported by theoretical considerations. Within this framework, we

conclude that a conformation will be scored more favorably than a second one if

- the frequencies with which the couples of types of residues are observed in a certain contact class is more similar to the ones observed in the learning data set. This guarantees that the right types of residues are making the right types of contacts;

- there is more "segregation" between types of contacts and associated couples of types of residues, i.e. the classes of contact select more specifically the types of residues involved. The condition holds unless it violates the previous one.

To have an idea of how these frequency distributions look like, in Fig. 4.10 we report the ones relative to the TOP500 dataset. The polar, apolar and non-contact parameters are visually compared. The residues are disposed along the axes in order of decreasing hydrophobicity. We can see the primary information contained in the apolar parameters is about polarity. In the polar parameters, we can observe that specific kinds of contacts have very high or very low values, thus reducing the entropy of the contact distribution. We also note that the two frequency distributions are uncorrelated. This proves that they do not contain the same kind of information. Unsurprisingly, the non-contact class is the least informative: its role is mainly enforcing normalization (see Section 4.2). Yet, there are at least three specific couples of types of amino acids for which this class of contact contains a sharp information: the residues CYS, PHE and TRP, when present, are very disfavored if not paired. We can explain the first one for the propensity of CYS to make sulfur bridges, and the other two because of their size and their propensity to form pi stacks.

4.6.2 A unified scoring function for protein folding and protein-protein interaction

We performed tests on decoy sets from the CASP [CKT09] and CAPRI [JHM$^+$03] competitions, showing that our scoring function performs better in recognizing the native conformation than some of the most popular state-of-the-art scoring functions in both problems. The performance is particularly remarkable in the docking context, since the relatively few parameters entering our scoring function are derived from TOP500, a dataset of monomeric proteins.

Further, the parameters weighting the relative contribution of different energy terms are not optimized. The remarkable portability of our potential is an indirect evidence that protein complexes are stabilized by the same fundamental interactions of monomeric proteins, as already shown in a similar way within the context of amyloid fibril aggregation [TCMS06]. In order to deal with the highly inhomogeneous nature of CAPRI decoy sets, we introduced a way to rank the native complex by comparing pairs of curated structures that share the same residue set. We suggest that the generalized ranking procedure introduced in this thesis could be useful for comparing inhomogeneous structures in general.
The success of the features that were introduced in BACH-S highlights the stabilizing role of the 'Janus face' ability of polar side chains in establishing favorable interactions with both their polar and non-polar moieties for both single chain proteins and protein complexes. Proper discrimination of unrealistic steric clashes in computational models is likewise crucial in both contexts. Our results suggest that the availability of higher quality decoy sets, with homogeneous models, refined and purified from bad steric clashes, is highly desirable for a proper benchmark of the different methods.
The importance of the contribution to the BACH-S score coming from residues that are distant from the interface underlines the role of the rearrangement of those residues in complex formation. From a practical perspective, a scoring function able to weigh on the same ground the contacts formed at the complex interface and the structural rearrangements occurring far from it may prove useful for addressing problems in which induced fit or allosteric regulation of the binders are crucial in determining the correct docking pose. These tasks are not conceivable for scoring functions that evaluate only the interface. Furthermore, we proved that by only considering the interface, the discrimination power of the statistical method weakens significantly.

4.6.3 Protein-protein interaction is still a challenge

In the more general perspective of protein-protein structure prediction, our analysis underlines once more that this task is extremely challenging for computational approaches. Even if our scoring function is capable of discriminating the correct structure, no significant correlation between the score and the interface RMSD with the experimental structures is visible. For most of the

decoy sets the Pearson correlation between the iRMSD and the score is below 0.3. Even more importantly, structures with low score seem to be present at any iRMSD, making the value of the score a poor guide in a structural search carried out in the whole conformational space. A possible interpretation of these results is that some important ingredient is missing in our scoring function. However, all the scoring functions we analyzed in this thesis show a behavior that is qualitatively similar to the one observed for ours. In particular, in all cases several structures with low score are present at high value of iRMSD (see Figures 4.8 and 4.9. This is a remarkable difference with respect to what is observed in decoy sets of monomeric proteins, where all the scoring functions, including ours, show a significant correlation between the score and the RMSD towards the native structure [CGL$^+$12].

If the behavior observed above is at least qualitatively correct, one has to conclude that the free energy landscape for protein-protein interaction does not possess a clear funnel, at least for large iRMSD. While the protein folding process, to take place efficiently, must follow an energy-guided pathway towards the native state, one can argue that protein-protein pairs can meet each other by simple diffusion, making the need of a funnel-shaped landscape biochemically less stringent. Under this hypothesis, the attempt of correlating a scoring function with the quality of a non-native structural model might be misleading. Indeed, decoy complex structures with low score and high iRMSD might signal physically viable docking poses that may transiently occur in the complex formation process [YPVV08]. The performance of BACH-S in model quality assessment makes it a reliable tool to refine the structural search in the close proximity of the native pose. BACH-S exhibits an excellent ability in recognizing high quality structures that are very similar to the native pose when a small fraction of best ranked poses is selected. This is benchmarked by all the three tests we illustrate in the present thesis: recognition of the native pose, fraction enrichment and lowest iRMSD within the best poses. The performance of BACH-S is particularly good in selecting very few best-ranked poses. Several high quality structures may be missed, but the true positives found by BACH-S are more numerous and of higher quality than for the competing scoring functions considered in this thesis. FireDock performs very well in finding structures with the lowest iRMSD within the best poses, but its fraction enrichment is significantly lower than the one of BACH-S when a fraction of up to the 5% of the best-ranked poses is selected, and lower than the one of

PIE*PISA above this threshold. PIE*PISA displays the best performance in the high sensitivity/low specificity setting implied by selecting a large fraction of best-ranked poses, when many more false positives are present. Moreover, PIE*PISA is also more reliable when similarity is present, but less marked.
These results are possibly related to the way parameters are derived in the two scoring functions: while BACH-S parameters are obtained from a set of crystallographic structures, PIE*PISA is trained on misfolded structures. Thus, the former scoring function performs better at establishing whether a structure is correct (native or very near native), whereas the latter is more suitable in recognizing to which extent a structure is incorrect. Interestingly, an optimal performance in the high specificity setting is displayed by PASTA, in the quite different context of protein aggregation prediction. PASTA is an algorithm based on a scoring function that adopts essentially the same parameters as BACH-S for β-bridge residue-residue pairwise contact classes [TCMS06, WSTT14]. It may be possible that this behavior is due to deriving parameters only from a data set of native conformations, as in BACH-S, without boosting the high sensitivity performance by explicit training on a data set of competing conformations, as in PIE*PISA.
We conclude with an example of how the ability of BACH-S in specifically detecting high-quality structures can be fruitfully used in a realistic refinement procedure. Molecular dynamics runs could be performed on a small set of putative high quality structures pre-selected among the many possible poses generated by docking algorithms. We will apply these ideas in the next chapter, where we will follow a refinement procedure and we will assess the performance of BACH-S, PIE/PISA and Rosetta. The efficiency and the reliability of the approach crucially depends on how close these starting conformations actually are to the native state; in that respect, BACH-S appears a very promising tool.

Chapter 5

Recognizing the correct structure of a protein-protein complex

Recognizing the native structure of a complex among a set of decoy structures does not solve the problem of protein-protein interaction, in the same way as discriminating the native fold of a protein among a set of decoys does not solve the problem of protein folding. The standard procedure to tackle both issues can indeed be divided into two separate steps: first, a set of plausible conformations must be generated. Then, the best prediction must be selected out of the set. The two parts are both crucial, and the quality of the results obtained in the second part depends greatly on the quality of the conformational search performed in the first one.

In Chapter 4 we focused on the second task, and compared the performances of some state-of-the-art scoring functions in discriminating the correct pose in a series of challenging decoy sets. This is indeed the moment in the procedure where a scoring function is more needed. Nonetheless, the ability of discriminating the native state is not, by itself, sufficient: if only the correct state is discriminated, while the almost-correct conformations are completely ignored, the practical use of the scoring is compromised. To be useful, a scoring function should reproduce at least on a local scale the shape of the free energy funnel around the native state. This implies that a correlation should be visible between the score assigned to a conformation and its proximity to the native state. As a direct continuation of the study described in Chapter 4, we hereby test our scoring function BACH and the competitors PIE/PISA and Rosetta along a complete docking procedure. The aim here is not to compare the performances of sampling algorithms, but to characterize the correlation

between the quality of the poses and the discriminatory power of the scoring functions. To generate the poses we will make use of a rigid docking algorithm. The poses will then undergo different cycles of refinement through more and more accurate molecular dynamics simulations.

As we discussed in Chapter 2, rigid docking [KKSE$^+$92, MRP$^+$01, WN00, PKWM00, GWA01] does not change the internal structure of the two subunits, and spans the 6-dimensional conformational space given by the relative rotations of one monomer around the other. Restricting to this space, one can obtain a dense and uniform scan of all possible poses. Nonetheless, the majority of monomers modify their internal structure to optimize the interaction with their partner. Therefore, flexible docking methodologies [BRW95, ESDW$^+$08, DNRB94, MR05] are often more accurate. Being the conformational space too large, flexible docking schemes cannot scan the entire space, but can be applied to the study of a suitably chosen subset.
In this study, our first step will be performed by a rigid docking algorithm, for two reasons:

- To avoid or limit biases given by internal scoring functions of the docking algorithm. Since the focus of the study will be on the predictive power of protein-protein interaction scoring functions, it is important to limit the influence of any other filtering method. However, as we explained, often it is not possible to completely remove the bias: even rigid docking programs make use of minimal filtering methods which are intrinsic to the algorithm. For example, all FFT methods include an evaluation through convolution in order to retain only one configuration among the set of configurations generated by shifts in a fixed direction.

- To sample the whole accessible conformational space. One of the main problems of flexible docking methods is that they risk to completely ignore the native state conformation if the site is not recognized among the plausible ones in the filtering process. In order to provide for a better test for the scoring functions, we instead wanted to retain as many different conformations as possible, and let the scoring functions act as filters.

Therefore, in order to generate a dense set of poses we chose the algorithm ZDOCK 3.0 [PHW11], already described in Chapter 2. ZDOCK generates 54000 poses of a dimer by rotating and translating the ligand around the receptor.

In doing this, it introduces only one source of unevenness, which is inherent to the FFT-based method. We will subsequently refine the conformations generated by docking by using molecular dynamics simulations. In particular, we will first use MD in vacuum to allow for relaxation of the internal degrees of freedom of the two subunits. Then, we will concentrate on a subset of the conformations that will be chosen with the help of the scoring functions used. The conformations in this subset will undergo a more refined and longer MD dynamics in explicit water. Throughout the whole procedure, we will analyze the conformations with the selected scoring functions. This will provide us hints about both the quality of the docking poses and the characteristics of the algorithms used to score them. We will analyze in detail the performances of some state-of-the-art scoring functions, namely BACH [SZC+13], PIE/PISA [VRE13] and Rosetta [RSMB04], described in Chapter 2, on decoys of a CAPRI Target.

5.1 Methods

5.1.1 Choice of the target

Among the targets proposed in CAPRI, we choose the dimer (PDB code: 1syx) composed by the spliceosomal U5 snRNP-specific 15 kDa protein (PDB code: 1qgv) and the CD2 antigen cytoplasmic tail-binding protein 2 (PDB code: 1gyf). The dimer was the object of CAPRI Target 22, belonging to Round 8. The round was subsequently canceled because of the leak of unauthorized pictures of the dimer on the internet before the conclusion of the competition, but the data about the predictions are available on the site. Possibly due to the cancellation of the competition, the total amount of predictions available is fairly small (77), and they are of poor quality: according to CAPRI criteria, only 4 out of 77 complexes are "acceptable, while all the others are "incorrect. Target 22 is also part of the test set we used to assess the quality of different scoring functions in a previous work [SGS+15] and described in Chapter 4. Because of the characteristics of the decoy set, it is quite easy for a scoring function to discriminate from the rest of the structures the native conformation, if the latter is added to the decoy set. Indeed, 4 scoring functions (BACH-SixthSense, PISA/PIE, IRAD, HADDOCK) out of the 6 we analyzed ranked the native in the top 10%. For the same reason, finding the structure which is nearest to the exact conformation is a hard task. The only scoring function that,

for the decoys at low iRMSD from the native, exhibits a Pearson correlation coefficient higher than 0.1 between the iRMSD and the score is FireDock, which in contrast is the worst in detecting the native state when it is added to the prediction set.

5.1.2 Choice of the scoring functions

We will consider three scoring functions already tested in the previous study (Chapter 4).

- BACH is the scoring function described throughout all this work. We will consider here the version described in Chapter 4 and previously denominated BACH-SixthSense.
- PIE/PISA obtained good results in the tests described in [SGS+15] and reported in Chapter 4. It is the only scoring function among those considered whose performance competes with the one of BACH. For a more detailed description of the algorithm, see Chapter 2.
- Rosetta did not perform as well as BACH and PIE/PISA in our tests, but obtained good results in true CAPRI competitions and is the reference method for the whole class of scoring functions to which the other two methods belong. The algorithm is based on the same statistical premises, but the functional form differ greatly from the ones used in BACH and PIE/PISA, thus providing an even more interesting comparison. We use Rosetta with the same settings employed in the previous chapter: Rosetta 3.4 with the "standard" set of score12 weights.

5.1.3 Quantifying performances

In order to assess the quality of the predictions of the scoring functions we made use of the same quantities described in Chapter 4: our main tool will be the correlation plot between the interface root mean square deviation (iRMSD) from the crystallographic native state and the value of the scoring function. The iRMSD is the root mean square distance of the C_α atoms placed at the interface in the native conformation. These are defined as those C_αs whose amino acid has at least one atom closer than 10 Å from an atom of the other subunit [Jan10].
The second manner of assessing the performance is the fraction enrichment

[WFLS04], already presented in Chapter 4. The fraction enrichment is based on ranking the conformations from the lowest to the highest and then computing the number of near-native conformations found in the first n% of the rank. Differently from our previous studies, we hereby consider as near-native the poses with an iRMSD smaller than 5 Å. Albeit this threshold may seem exceedingly high, less than 1% of the poses generated via rigid docking respect the constraint. We remark that the correct threshold for this analysis is dependent on the size of the complex.

5.1.4 Generating a large set of rigid poses

The first step of our analysis procedure consists in generating a large set of rigidly docked binding poses. At this scope we employed the well-established docking algorithm ZDOCK 3.0 [PHW11] whose performances have been tested throughout the years in all CAPRI competitions. See Chapter 1 for a description of the competition. We generate 54000 conformations, which is the maximum possible number of poses that ZDOCK can provide. This is done in order not to make ZDOCK apply any additional filter to the already halved number of conformations and to span the largest possible fraction of the conformational space given by the three Euler angles (θ, ϕ, ψ).

5.1.5 Equilibrating the binding poses by short MD simulations in vacuum

The structures generated by ZDOCK are then equilibrated by performing 40 ps of plain MD in vacuum. We use GROMACS 4.6.7 [HKVdSL08] to perform the simulations. Since the receptor has a hole of six amino acids in the sequence, we keep the position of the two C_αs next to the hole fixed. We rescale the masses of every atom to be equal to 2 Da. This allows exploring more conformations per unit time, but does not affect thermodynamics quantities. We run the simulation at a temperature of 300 K, and with a cutoff for the electrostatic interactions of 1.2 nm. Of the initial 54000 conformations, 39286 complete the molecular dynamics simulation. The conformations that do not complete the dynamics are normally characterized by serious steric clashes between the subunits, that our computational protocol is not able to cure. This conformations are excluded from any further analysis. The conformations that complete this step are then scored by three different scoring functions, as we

will discuss in detail in the next section.

5.1.6 Equilibrating the system by MD in water solution

For all the poses reaching the end of the equilibration in vacuum we split the trajectory into four parts of 10 ps each and calculate the average BACH energy and its standard deviation on these four parts: E_B^i, σ_B^i, for $i = 1, 2, 3, 4$. We select only those trajectories that satisfy these conditions:

$$\begin{aligned} E_B^1 &> E_B^3 \\ E_B^2 &> E_B^4 \\ E_B^4 - E_B^3 &< \frac{\sigma_B^3 + \sigma_B^4}{2} \end{aligned} \tag{5.1}$$

In this manner we exclude from any further analysis the conformations whose score significantly increases during the first relaxation step. We therefore assume that a conformation that violates these conditions is unstable. After applying this filter, 34253 conformations remain. We then take the top 2000 structures satisfying this criterion and simulate them in explicit water for a total time of 1.3 ns. Again, we use GROMACS 4.6.7 [HKVdSL08], with a 2 fs time step. We freeze the C_αs of the receptor and we put the masses of the atoms to 2 Da, as in the previous step. Each system undergoes temperature and pressure equilibration phases of 100 ps and 200 ps respectively. We use a velocity-rescale algorithm as thermostat, Berendsen barostat [BPvG$^+$84] for the NPT equilibration and Parrinello-Rahman barostat [PR81] during the dynamics. The relaxation time for the velocity-rescale thermostat is 0.1 ps, the one for the Berendsen barostat is 2 ps, and the one for the Parrinello-Rahman is 2 ps. The cutoff for electrostatic interactions is 1 nm.
1990 of the simulated structures complete the dynamics. Some of the simulations are then continued for other 100 ns in order to perform additional analyses on the stability of the structures.

5.1.7 Scoring the crystallographic structure

The docking procedure is carried out as if the native structure was unknown. Yet, we use the crystallographic structure of the dimer as a reference to evaluate the quality of the models. As we proceed through the different steps, we process the PDB dimer structure as well, in order to produce a consistent score of

for the native pose at each stage. This additional analysis will also provide information about the stability of the crystallographic conformation, thus providing a benchmark of the estimators we are using.

5.2 Results

The analysis we present in this work is aimed at benchmarking the separate importance in a protein-protein prediction problem of two elements: the quality of the scoring function and the quality of the structures that are scored. For performing the analysis we choose a single target from the CAPRI competition, the dimer of PDB code 1syx composed by the spliceosomal U5 snRNP-specific protein (PDB code: 1qgv) and the CD2 antigen cytoplasmic tail-binding protein 2 (PDB code: 1gyf). This target was selected because the size of the monomers is not too small and not too large (130 and 62 residues), and because the quality of the predictions available in CAPRI for this case is rather poor (see previous section).

We evaluate the quality of the structures with three different scoring functions: BACH-SixthSense, PISA/PIE and Rosetta. BACH-SixthSense was proved to have the best discrimination power among other scoring functions (PISA/PIE, IRAD, Rosetta, FireDock, HADDOCK) when near-native states are present, while for more distant conformations it is overcome by PISA/PIE. Rosetta is an extremely popular protein/protein interaction tool, whose usefulness was demonstrated several times even in blind predictions [FCS+10].

The main scope of this work is benchmarking the relative importance of the quality of the scoring functions and the quality of the structures. Since different algorithms generate high quality structures with a different efficiency, we here on purpose to avoid generating structures by any advanced importance sampling technique as those implemented in Rosetta [WSFB05] and HADDOCK [DBB03]. We instead perform the analysis starting from all the conformations that are geometrically possible, generated by a rigid docking algorithm. The first refinement step is performed "blindly" on all these structures. The subsequent step is based on costly atomistic simulations in explicit solvent. In order to limit the computational burden, the data set was reduced by selecting the best poses according to BACH-SixthSense. However, in order to avoid as much as possible biases induced by the selection procedure, we performed the refinement step on the largest number of structures we could afford with our computational

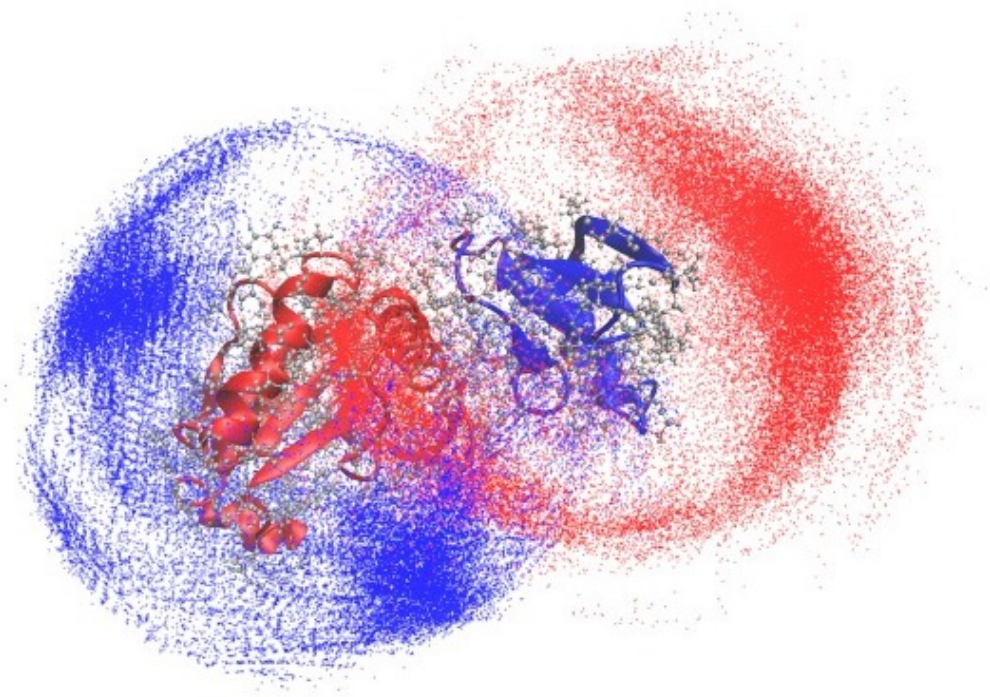

Figure 5.1: Native conformation surrounded by the cloud of the centers of mass of the 54000 generated conformations. The red dots and blue dots represent the centers of mass of the receptor around the ligand and of the ligand around the receptor, respectively.

resources.

5.2.1 Step 1: analysis of the poses generated by rigid docking

The first step of our analysis procedure consists in generating a large set of rigidly docked binding poses by ZDOCK 3.0 (see Section 2). Figure Fig 5.1 shows the position of the centers of mass of the ligand around the receptor and viceversa, for each of the 54000 poses. Some density holes can be found in both clouds, showing that ZDOCK 3.0 depopulates some areas and favors others. This is due to the selection procedure described in Chapter 4. One of the holes is spatially close to the native state configuration, yet not enough to compromise the search: at his stage, 257 near-native poses are present in the set.

Fig. 5.2 reports the iRMSD as a function of the score of the three scoring functions, for each one of the refinement stages considered. The first row of the graph represents the initial situation. We show that at this stage none of the

three scoring functions considered display any visible correlation between the two quantities: the near-native poses are randomly placed halfway through the rank. The graph also reports the score of the native conformation, visualized in each panel as a black diamond. We can see that BACH and Rosetta correctly discriminate the crystallographic native state at this stage, while PIE/PISA assigns to it a lower score than the near-native conformations but a higher score than some other decoy. A certain similarity in the shape of BACH and Rosetta point clouds is also noticeable. In fact, the two scores display a Pearson correlation coefficient of 0.87. We will see that the correlation will be lost as the refinement proceeds.

5.2.2 Step 2: analysis of the poses after structural relaxation in vacuum

For each of the 54000 poses we than performed a 40ps dynamics in vacuum, with the scope of relaxing the structures and curing the steric clashes unavoidably generated by rigid docking. 39286 simulations complete this step without errors, while the remaining exploded due to too extreme steric clashes (see previous section). The second row of Fig. 5.2 shows the iRMSD versus the average value of the scoring function computed on the last 10 ps of these simulations. At this stage, Rosetta is the only scoring function which discriminates as top-ranked a near-native state. We will see, however, that when the structures are further refined this does not happen anymore. In all the scoring functions, at low iRMSDs a very mild correlation starts to emerge. This change highlighted by the fraction enrichment estimator represented in the first row of Fig. 5.3. In each panel of the first row, the solid line represents the fraction enrichment during Step 1 and the dashed line the fraction enrichment during Step 2. The fraction enrichment for the configurations after the 40 ps dynamics grows more rapidly, thus showing that a larger number of near native poses can now be found in the high part of the rank. The total amount of poses and the total amount of near-native states differ in the two stages: originally, 257 out of 54000 initial poses are near-natives, while after the dynamics the near-natives are 192 out of 39286 remaining poses. The correlation between the scores of BACH and Rosetta completely fades during this stage, passing from a Pearson correlation coefficient of 0.87 in the previous step to a value of 0.07 at the end of this step.

The analyses show that, even after the MD relaxation, two out of three scoring

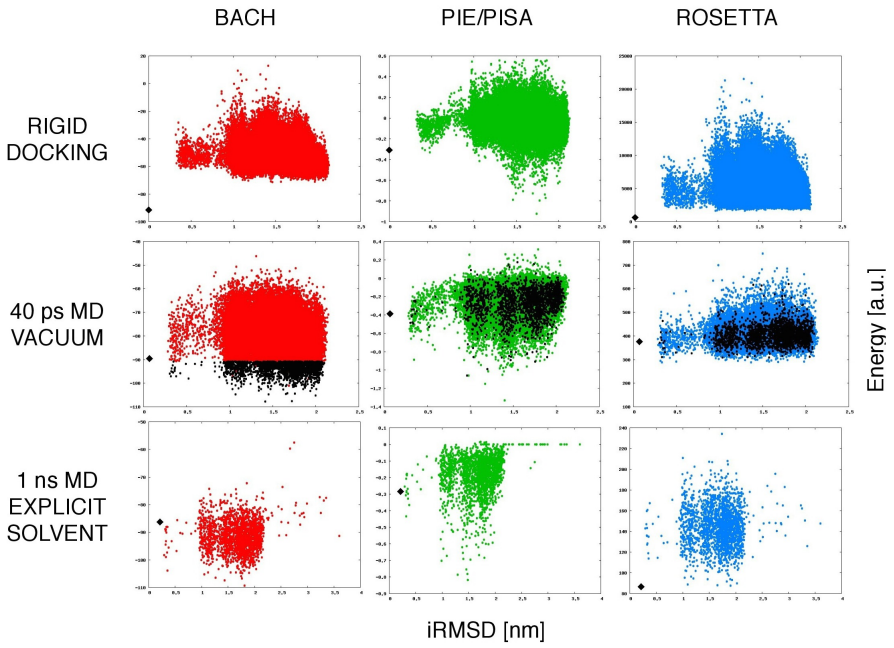

Figure 5.2: iRMSD vs. the scoring functions in the different steps of the procedure: a) for the 54000 generated decoys, b) for the 39286 decoys after the equilibration in vacuum, c) for the 1990 best-scored decoys after the equilibration in explicit water. In all panels is reported the score of the native conformation (marked in black), which underwent the same procedure as the decoys.

functions cannot predict the correct conformation of the dimer. Indeed, the best near-native pose is ranked 222 by BACH and 36 by PIE/PISA. Yet, the success of Rosetta and the emerging correlation are hints that MD could help the discrimination process. Moreover, in all three cases the fraction enrichment is significantly better after the MD relaxation than before. The natural next step to do is to increase the accuracy of the simulations.

5.2.3 Step 3: analysis of the top-2000 poses by a 1 ns MD in explicit water

We then selected the top-2000 poses with the lowest BACH score from the previous step (see Section 5.2.2). The chosen poses are colored in black in the second row of Fig. 5.2. We can see that, due to the very low correlation between the scores of the three scoring functions, the poses are randomly distributed in the ranks of PIE/PISA and Rosetta. The choice of these 2000 poses constitutes a source of bias in the comparison of the performances of the scoring functions. From now on, we will therefore only discuss how the refinement steps impact the

behavior of the scoring functions, without comparing their relative performances. The bias is mitigated if we keep in mind the properties of the distributions of the chosen points during Step 2 (the black clouds in Fig. 5.2). Among the selected poses, PIE/PISA ranks the first near-native as 172th, and Rosetta as 5th.

Each structure was solvated in a box containing on average 10400 water molecules and equilibrated at room temperature and pressure for 0.3 ns, followed by 1 ns of production (see Section 5.1.6). Since this procedure is computationally expensive, it was not possible to select the top-2000 poses for each scoring functions independently, and we carried on this procedure only for BACH. A molecular dynamics of 1 ns in explicit solvent contributes to thermalize and relax the originally coarse structures, stabilizing the contacts that are chemically viable, and breaking the ones that are stable only in vacuum. Comparing the second and the third row of Fig. 5.2 we see that while for BACH there is no qualitative change, for PIE/PISA and for Rosetta there are substantial differences. PIE/PISA now ranks the first near-native conformation as 77th, thus improving the initial situation. However, the PIE/PISA panel in the third row of Fig. 5.2 shows that the fraction of low-ranked wrong predictions increased. Regarding Rosetta, we notice that it is not able to rank in the first places a near-native conformation anymore, placing the first one as 90th. Also in this case, some of the wrong predictions got a much lower score. The discrimination power of BACH remains qualitatively the same, even if now the first near-native can be found as 16th in the rank. All the three scoring functions still exhibit a mild correlation between iRMSD and score at low values of iRMSD. We remark that in this section we are discussing the impact on the rank of the refinement step on a set that includes the same structures but is optimal only for BACH.

The analysis of the fraction enrichment is reported in the second row of Fig. 5.3. The dashed lines here represent the fraction enrichment during Step 2, but only limited to the selection of 2000 structures that will then undergo the dynamics in explicit solvent. The dotted lines represent instead the fraction enrichment during Step 3, after the dynamics in explicit solvent. This analysis reveals that in the quantity of top-ranked near-natives slightly increases for BACH and PIE/PISA, while greatly decreases for Rosetta, for the same reasons we stated during the analysis of the ranking.

Particular attention should also be given to the position of the native structure

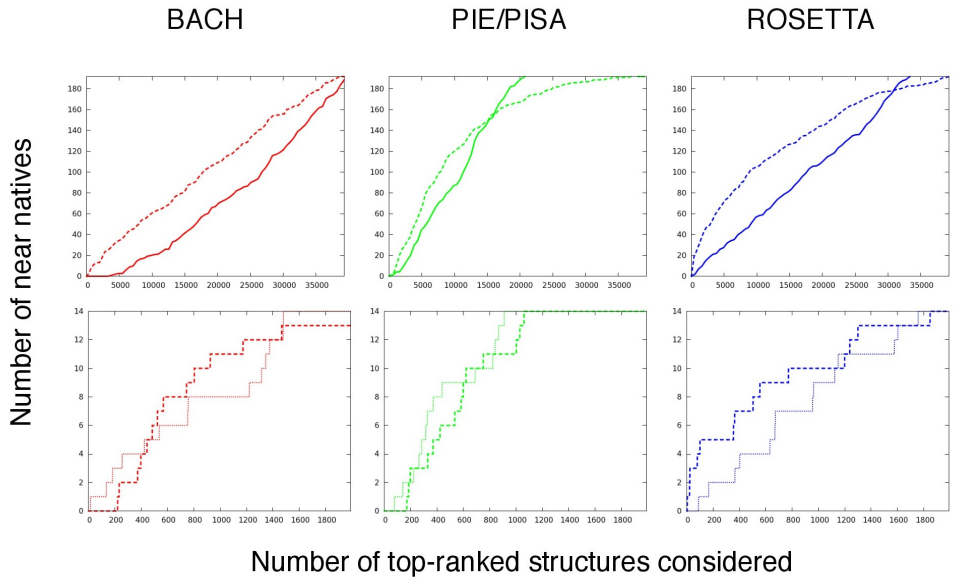

Figure 5.3: Fraction enrichment for the three scoring functions at different stages of the procedure: in the first row the fraction enrichment relative to the rigid poses of Step 1 (solid line) is compared with the fraction enrichment relative to the poses relaxed with a 40 ps MD in vacuum of Step 2 (dashed line). In the second row, the enrichment factor relative only to the 2000 poses selected by BACH is reported after that the structures underwent the MD in vacuum of Step (dashed line) and after also the MD in explicit solvent is performed in Step 3 (dotted line). We consider as "near native" those conformations having iRMSD < 0.5 nm

in the graph. The iRMSD of the simulation started from the native state (black diamond) is drifting towards higher values, while this is not happening for some of the near-native conformations. This implies that the equilibrium conformation is actually at more than 2 Å of iRMSD from the alleged native state, and that not even the simulation which takes as starting configuration the crystallographic structure has reached convergence yet.

5.2.4 Scoring by 100 ns of MD in explicit water

The previous step is computationally expensive, since it is based on running a total of 2.6 μs of molecular dynamics in explicit solvent. However, this procedure revealed not yet sufficient to find the native structure. Apparently, the poor quality of the results does not depend on the specific scoring function, but is primarily determined by the quality of the configurations. Indeed, the predicting power of BACH and PIE/PISA scoring functions increases whenever the conformations become more and more realistic. Remarkably, this happens

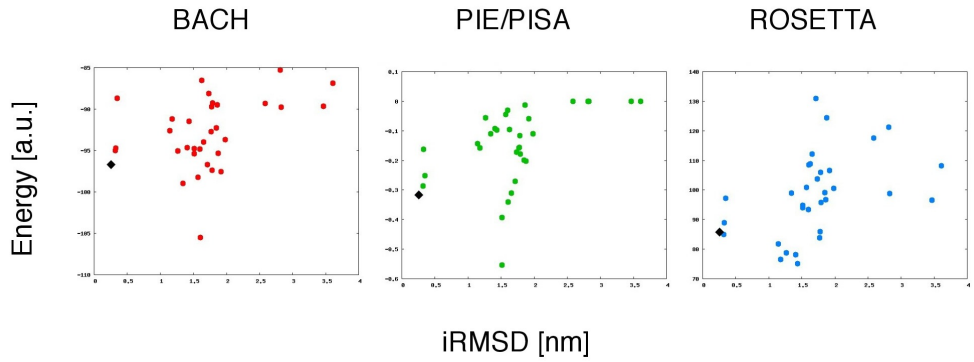

Figure 5.4: iRMSD vs. the scoring functions for 33 selected decoys which underwent a 100 ns dynamics in explicit solvent. The score is averaged on the final 50 ns of the dynamics. In all panels is reported the score of the native conformation (marked in black), which underwent the same procedure as the decoys.

also for the crystallographic structure, that is scored relatively badly by BACH, indicating that it is at least partially distorted by crystal packing effects, and is therefore not fully representative of the thermal ensemble in water solution. In order to investigate how important this effect is we extend the dynamics of some representative conformations up to 100 ns. We choose to perform this additional dynamics on a restricted set of near-native states and the far-from-native conformations associated with the lowest energy scores. We simulate the native conformation as well. It is found that 100 ns allow most conformations to reach stability. Except for few cases, after 50 ns the score as a function of time reaches a plateau. Thus, the score can be estimated as an average on the second half of the dynamics. Fig. 5.4 shows the final situation: in all three scoring functions the native and near-native states have similar scores. Their rank improved but is not yet optimal, indeed in all three cases there are structures at higher iRMSD that display lower scores.

5.3 Discussion

In this study we compared the performance of different state-of-the-art scoring functions for docking pose prediction on a binary complex already proposed in CAPRI as Target 22. Starting from a very large set of poses generated by the rigid docking algorithm ZDOCK 3.0, we performed molecular dynamics with an increasing degree of accuracy. At each step, we assessed the discrimination proficiency of some scoring functions in recognizing the native pose. For

BACH and PIE/PISA the predictive power increases with the accuracy of the simulation, while for Rosetta it decreases after the 1 ns optimization in explicit solvent. After 100 ns of dynamics in explicit water the predictive power of all three scoring function increases, but the native structure cannot be discriminated from all the decoys yet.

5.3.1 Generation of rigid poses

We chose to rely on a rigid docking procedure for two reasons: first, this approach allows generating a large set of decoys which span most of the rigid conformational space, thus providing a rough but complete overview of the possible docking poses. Second, in this way we avoid the biases that a flexible docking method implies. However, we did not succeed in removing all the biases from the docking algorithm, as we can see from Fig 5.1. There is indeed one type of bias which is intrinsic to the use of Fast Fourier Transforms. For two average-sized monomers, a uniform and fine conformational search in the 6-dimensional space of translations and rotations would imply the generation of hundreds of billions of poses (see for example [MRP$^+$01]), most of which would be completely unphysical. Thus, the current methods which rely on such a search usually perform explicit rotations and then use FFTs and the convolution theorem to compute at once the score of all poses separated only by a translation in a same direction. As we explained in Section 2.4, this implies that poses in which the ligand happens to be on opposite sides of the receptor fall in the same set of translations. There is no way to ensure that both configurations are recognized: if the algorithm retains only the best scored pose for each set of translations, one of them is lost. If the algorithm retains more than one, still there is no guarantee, because it could (and possibly does) retain only configurations of the ligand on the same side of the protein, where the scoring function is more favorable. With an angular step of 6 degrees, 54000 poses are generated. If the translational shifts are 1 Å long, to be somehow sure to retain at least one pose of the ligand on each side of the protein one should retain on average 10 structures for each translational set, which would imply to have an output of 540000 poses. The current analysis would not be feasible with such a number of candidates. Although we are aware that specific methods could be devised in order to select the two poses, the effort demands a study by itself, and falls outside the scope of our work. The choice of the rigid docking algorithm fell on ZDOCK because it provides the best compromise

between refinement of the search and quality of the selection. DOT [RTP+13] was a good alternative, but produces a larger number of candidates without eliminating the source of bias described above.

5.3.2 Performance of the scoring functions through the different refinement steps

We analyzed the performance of three scoring functions along the whole docking procedure. The results are summarized in Fig. 5.2 and 5.3, where we can see the iRMSD vs. score correlations and the fraction enrichments respectively, for each scoring function during each of the refinement steps. We already concluded that after the relaxation in vacuum the performances of all three scoring functions improve. In particular, during Step 2 Rosetta is able to discriminate as top-ranked structure a near-native pose. Although this is a very good result, the iRMSD vs. score plot for Rosetta reveals that there might be also a fortuitous component: indeed the distribution does not show features as outstanding as the ones of the rank. The correlation is mild and broad as in the other cases. While BACH and PIE/PISA are based on contact counting, Rosetta is based on more complex structural considerations. This can represent a hint on the fact that contacts are not the best feature to discriminate between semi-rigid poses, while they become more informative as the quality of the poses increases, as we will see hereafter.
After the stage of refinement in vacuum the procedure becomes biased by the selection of the best 2000 poses according to BACH. We can see from the second row of Fig. 5.2 that the selected poses are not among the best-scored ones for the other two scoring functions. Fig. 5.6 shows for each scoring function the score distributions for both the whole set of 39286 poses (solid line) and for the subset of 2000 selected poses (dotted line). The fourth panel reports the distribution of the iRMSD for the two sets considered. We can see that for both PIE/PISA and Rosetta the selection of the best-scored 2000 poses in BACH does not differ significantly from a random selection. The distributions relative to the iRMSD report differences in the height of the peaks: there is a slight increase in the fraction of near native poses (it is higher in the selection for Step 3) and there is a higher percentage of poses with iRMSD > 1.5 nm. The small peak at low iRMSDs is only a consequence of the improved BACH fraction enrichment profile: by taking the 2000 top-scored structures, there is a sligthly higher percentage of near-native structures than by taking 2000 randomly cho-

sen poses. Interestingly, it seems that BACH selects also structures with higher iRMSD. We suppose that this effect could be attributed to conformational entropy: in the case of a uniform sampling of the conformational space, the number of poses with iRMSD between a value v and $v + \Delta$ nm is lower than the number of poses with iRMSD between a higher value V and $V + \Delta$ ($v < V$), independently of the choice of the reference structure.
If these considerations prevent us from comparing the performance of BACH with the other two scoring functions after Step 2, they do not prevent us from comparing the performance of each single scoring function throughout the whole procedure. Fig. 5.3 shows that the dynamics in explicit solvent favors BACH discriminative power the most, while also favoring the one of PIE/PISA. On the other hand, Rosetta greatly loses the efficiency shown during the previous step. This can also be seen in Fig. 5.5, where the fraction enrichments already shown in Fig. 5.3 are reported together for Step 2 and Step 3, allowing a visual comparison of the performances of the three methods. We can see that, after the 1 ns dynamics in explicit solvent are performed, for low values of the fraction enrichment BACH is the best scoring function, while its performances are overcome by PIE/PISA in the rest of the graph. Rosetta, which was by far the most performing scoring function after the simulation in vacuum, attains now the worse results. This suggests that, as the quality of the structure proceeds, the information stored in the contact map and used by BACH and PIE/PISA becomes more and more useful, while the one provided by Rosetta seems less reliable

5.3.3 The native state differs from the crystallographic structure

Another important observation to make on the data shown in Fig. 5.2 is that the native state displays a significant variability both in iRMSD from the crystallographic structure (i.e., from its starting configuration) and in the score, for all the scoring functions considered. At the end of the 100 ns dynamics in explicit solvent, the structure seems to have reached stability at approximately 2.5 Å of iRMSD from the crystallographic structure. For all three scoring functions, the final score of the native state is lower than most of the other structures that also underwent the 100 ns dynamics, and is comparable to the score of the near-natives. However, the score of the native state during the

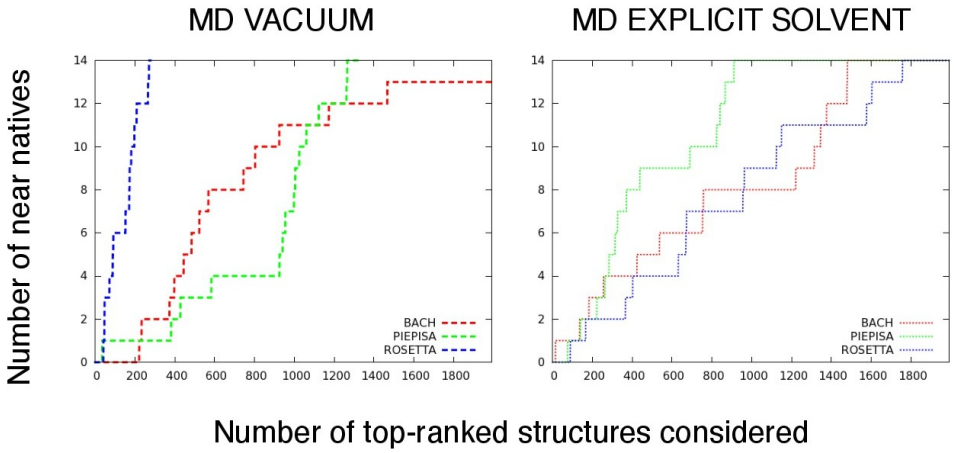

Figure 5.5: Comparison of the fraction enrichment for the three scoring functions. The analysis is restricted to the 2000 poses selected by BACH. The two panels summarize two different stages: after that the structures underwent the MD in vacuum of Step 2 and after also the MD in explicit solvent is performed in Step 3.

first 1 ns dynamics in vacuum shows a significant rise in BACH and PIE/PISA, while Rosetta displays a constant decrease of the native score throughout the whole procedure.

5.3.4 Characteristics of the contact features

The previous observations indicate that there is a substantial difference between the predictive power of BACH and PIE/PISA on one side and Rosetta on the other side. The 1 ns dynamics in vacuum is shown to enhance the predictive power of BACH and PIE/PISA. This is expected, since allowing for more flexibility helps the structure to recover the favorable contacts if the pose is stable, and to be driven adrift by the entropy if the conformation is unstable. However, we observe that Rosetta reacts differently, by losing almost all its predictive power, which instead was found to be the highest for structures relaxed in vacuum. While BACH and PIE/PISA are only based on the observation of contacts, Rosetta also grasps different structural features such as rotamers probability and secondary structure elements, from which orthogonal information can be extracted. This additional information seems to improve the performance of Rosetta when the structures are similar to the crystallographic ones, but to make the scoring function lose accuracy when the structures become more realistic due to the interaction with the solvent. In contrast, relying only on the contact map makes BACH and PIE/PISA less

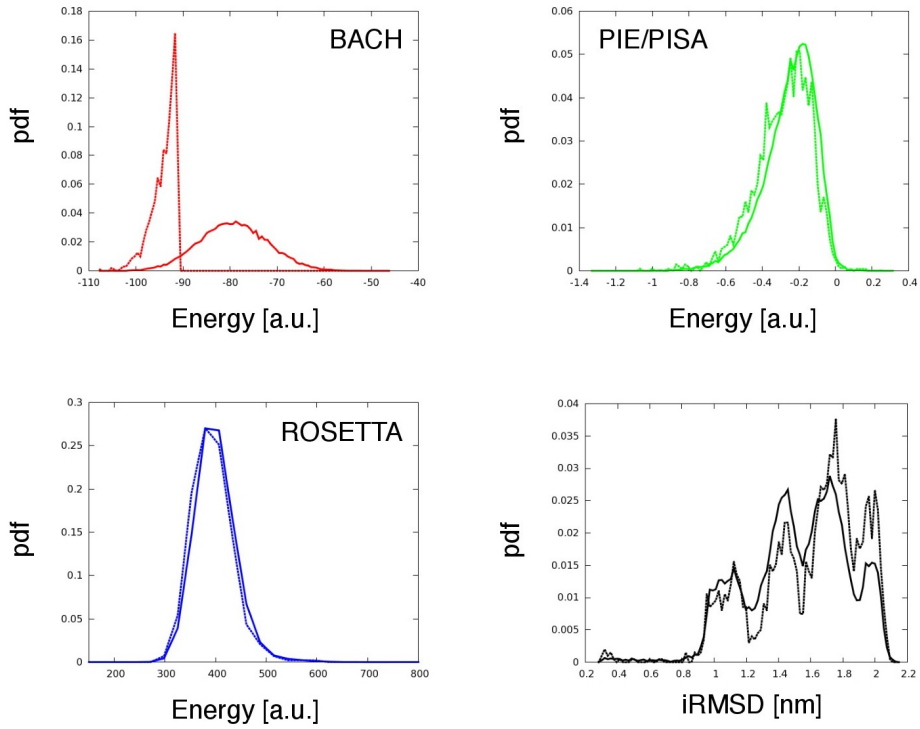

Figure 5.6: Probability density distributions of the three scoring functions BACH (red), PIE/PISA (green), Rosetta (blue) and of the iRMSD, both for the complete set of 39286 poses of Step 2 and for the selection of 2000 poses used in Step 3. Since the selection has been made by considering the 2000 top-ranked by BACH, the PDFs for that scoring function largely diverge. For the other two scoring function, the two PDFs display a marked similarity. The PDFs relative to iRMSD display noticeable differences, but the overall shape remains the same.

accurate when dealing with rigid structures, but subsequent refinement of the quality of the poses entails a steady improvement in the accuracy of the rank.

5.3.5 Docking and scoring as part of a single procedure

If methodologically it is convenient to separate the docking procedure and the refinement, it is not clear whether the division into two complementary tasks also leads to the best results [VHK13]. The results presented in this chapter brings two important messages:

- The quality of the docking poses largely affects the predictive power of the scoring function used for refinement
- The best scoring function at different refinement steps is not necessarily the same

These evidences suggest that current docking procedures are optimized when coupled to specifically devised scoring functions. This however implies that the predictive power of current docking methods does not depend on the realization of a general "physical fitness", rather on a specific fitness which is defined by internal criteria, which may or may not correspond to actual physical characteristics of the structure. We are confident that an objective set of features to correctly discriminate the native state of a protein complex from decoy states must exist. In our opinion, the quantity and quality of inter-residue contacts should be part of this information. The results show to support this hypothesis, though improvements in the quality of the information extracted from the contact map should be pursued.

Chapter 6

Conclusion

Protein folding and protein-protein interaction are two of the fundamental processes in cell biology. They are both central in the field of medical treatments: many neurological syndromes are caused by a disruptive folding of some proteins, while knowledge of the interaction mechanisms and the protein-protein interaction network in the cell are of key importance for pharmaceutical studies, indeed most of the currently used and developed drugs act by enhancing or preventing proteins to interact. This makes the two processes interesting to study by means of a physical approach.

The main goal of this thesis was to inspect the connection between the processes of protein folding and protein-protein interaction through computational methodologies. In particular, we made use of BACH, a scoring function conceived for protein folding problems. We proposed a series of improvements that made it suitable also for studying protein-protein interaction:

- The solvent interaction was studied by implementing a new calculation of the residue-wise solvent accessible surface area (SASA), in order to better discriminate between the residues which are exposed to the solvent and the ones which are buried in the core of the protein;
- The ability to recognize the correct conformation in protein-protein interaction problems was achieved by implementing in BACH a term taking into account steric clashes and a term describing the interactions involving sidechains with a partially hydrophilic and partially hydrophobic character;
- The ability to recognize the most correct pose among the ones calculated

by a single docking algorithm was tested by following a protocol involving docking and molecular dynamics at various levels of accuracy, using BACH and other two state-of-the-art scoring functions in each step of the refinement process.

The thesis can thus be divided into three parts, each of which focused on one of these three subjects.

Implementing the mLCPO method to account for the residue-wise exposure to the solvent We devised and tested a simplified version of the LCPO method [WSS99] to calculate the solvent exposed surface area of the residues of a protein. We call our version of the method *modified LCPO* (mLCPO). In order to produce an estimate of the solvent accessible surface area (SASA), the original LCPO method uses four parameters for each atom element and type of hybridization. Instead, we used the same set of four parameters for every heavy atom. We then promoted the water probe radius and the buried/exposed threshold on the SASA value to free parameters, and used them to optimize the coherence score of mLCPO with respect to two reference methods: SURF [VBW94], which is the method previously employed in the molecular surface area calculation in BACH, and GETAREA [FB97], an exact analytical method for SASA calculation. Although the maximum was quasi-degenerate, a meaningful value could be found for these parameters. We chose as optimal threshold and radius the values 0 and 3.08 Å, respectively. These values do not differ significantly from the ones reported in [SZC$^+$13]. The implementation of mLCPO conferred to BACH a speed 10 times larger than the original version, and a slightly better accuracy. Moreover, mLCPO allows to compute the derivatives of the SASA with respect to the atomic coordinates. This is important in view of an implementation of BACH as a collective variable in enhanced sampling methods.

Protein folding and protein-protein interaction by the same scoring function We devised a new version of the BACH scoring function [CGL$^+$12, SZC$^+$13] that can be used to evaluate protein conformations both for single protein chains and for protein-protein complexes. Two distinctive features of this version are the splitting of polar and apolar sidechain contacts into two different classes, and a potential energy term disfavoring steric clashes. Including these modifications is crucial in the protein-protein interaction problem, and

significantly improves the performance also for monomeric proteins. The BACH score introduced in this thesis is then

$$E_{\mathrm{BACH}} = p_1 E_{\mathrm{pair}} + E_{\mathrm{sol}} + p_2 E_{\mathrm{clash}} \tag{6.1}$$

where $p_1 = 0.6$ and $p_2 = 0.018$. p_1 was optimized in a previous work [CGL$^+$12], while p_2 was introduced and optimized in the work described in Chapter 4. The optimization followed the same criterion of the optimization of the weight p_1. We summarize here the three contributions to the total score:

- E_{pair} is the pairwise interaction term already contained in the original BACH algorithm [CGL$^+$12]. However, in the current version there are six types of contact between residues: parallel and anti-parallel β-sheet, α-helix, sidechain-sidechain Van der Waals polar and apolar contact, non-contact. The introduction of the separation between polar and apolar contact was justified by a theoretical point of view and was proved to enhance the performance of the algorithm both in protein folding and in protein-protein interaction problems.

- E_{sol} is the protein-solvent interaction term based on the new method to calculate the solvent exposed surface area described in Chapter 3. This conferred more speed to the algorithm and gave the possibility to calculate derivatives of the solvent exposed surface area with respect to the atomic coordinates.

- E_{clash} is a new term that accounts for steric clashes. Unlike the other ones, it is not a knowledge-based potential built on the concepts explained in Chapter 2. The clash term is atom-wise, and is equal to zero if two atoms of a certain element are further than a certain threshold. Otherwise, a disfavoring quadratic term is added to the score. The parameters of the quadratic formula depend on the element of the two atoms involved and are extracted from the probability distributions of the distances of couples of atoms in the proteins of the TOP500 database.

We performed tests on decoy sets from the CASP [CKT09] and CAPRI [JHM$^+$03] competitions, showing that our scoring function performs better in recognizing the native conformation than some of the most popular state-of-the-art scoring functions in both problems. In particular, on protein-protein interaction problems we tested BACH against five renowned scoring functions:

PIE/PISA, Rosetta, IRAD, HADDOCK and FireDock. The performance is particularly remarkable in the docking context, since the relatively few parameters entering our scoring function are derived from TOP500, a dataset of monomeric proteins. Furthermore, the parameters weighting the relative contribution of different energy terms are not optimized for protein-protein interaction. The remarkable portability of our potential is an indirect evidence that protein complexes are stabilized by the same fundamental interactions of monomeric proteins.

Recognizing the structure of a protein-protein complex To further validate BACH performance on protein-protein interaction problems, we compared the performance of different state-of-the-art scoring functions for docking pose prediction on a binary complex already proposed in CAPRI as Target 22. Starting from a very large set of poses generated by the rigid docking algorithm ZDOCK 3.0, we performed molecular dynamics with an increasing degree of accuracy. At each step, we assessed the discrimination proficiency of BACH, PIE/PISA and Rosetta scoring functions in recognizing the native pose. For BACH and PIE/PISA the predictive power increases with the accuracy of the simulation, while for Rosetta it decreases after the 1 ns optimization in explicit solvent. After 100 ns of dynamics in explicit water the performance of the three scoring functions increases, yet they still cannot discriminate the native structure from some of the decoys. This suggests that the quality of the docking poses largely affects the predictive power of the scoring function used for refinement.
Moreover, the best scoring function at different refinement steps is not necessarily the same: indeed, there is a substantial difference between the predictive power of BACH and PIE/PISA on one side and Rosetta on the other side. The 1 ns dynamics in vacuum is shown to enhance the predictive power of BACH and PIE/PISA, while we observe that Rosetta reacts differently, by losing almost all its predictive power, which instead was found to be the highest for structures relaxed in vacuum. This difference can in part be attributed to the different approaches by which the three scoring functions are built: while BACH and PIE/PISA are only based on the observation of contacts, Rosetta also grasps different structural features such as rotamers probability and secondary structure elements, from which orthogonal information can be extracted.
These results confirm that current algorithms for prediction the structure of protein-protein complexes are optimal when coupled to specifically devised

scoring functions. This however implies that the predictive power of current docking methods does not depend on the realization of a general "physical fitness", but rather on a specific fitness which is defined by internal criteria, which may or may not correspond to actual physical characteristics of the structure.

6.1 Future perspectives

The algorithm BACH for the discrimination of native and near-native conformations of proteins and protein complexes already proved to be quite reliable: it achieved top performances against other state-of-the-art scoring functions when tested on two collections of CASP and CAPRI decoy sets. However, especially for protein-protein interaction problems, the quality of the predictions of all the algorithms tested is still to be improved. It is indeed possible that BACH and other scoring functions lack some important ingredient. Here we list some of the ideas that could be developed.

- The possibility to add a term to account for the vibrational entropy change between the docked and undocked configurations of a dimer was explored, but up to now did not give positive results. Yet, we think that a better insight on entropic contributions might improve the current estimates of the free energy of binding.

- The ideas contained in the statistical approach presented in Chapter 4 can be further expanded and generalized, in order on one side to supply an original overview of the current knowledge-based potentials, and on the other to provide a theoretical basis for further improvements of the BACH scoring function.

- The introduction of system-specific terms accounting for coevolutionary data could improve the accuracy of the contact potential. Recent works [CT15, LT13, MPL$^+$11] show how patterns of co-evolving amino acids highlight important contacts in native structures. The inclusion of this kind of information in the theoretical framework delineated in Chapter 4 poses an interesting challenge, and goes in the direction of the upgrade from the noiseless to the noisy information formalism [Sha49].

Another direction that might be explored in the future is developing a BACH scoring function capable of predicting the interactions between proteins and

small ligands, a central matter in pharmaceutical studies. However, this implies to change from a residue-wise description to one which considers chemical functional groups as fundamental units. An early implementation can be already found in Zamuner's Ph.D. thesis [Zam15].
The advancements proposed in Chapter 4 also prelude to a further possible improvement: using BACH scoring function as a collective variable for enhanced sampling molecular dynamics. The implementation in Plumed [BBB+09] and Plumed2 [TBB+14] has been already completed, and preliminary tests on long molecular dynamics have already been performed.
We also considered using BACH as an energy term in a coarse grained force field. A first implementation in Gromacs 4.6.7 evidenced problems in the stability of the secondary structure and issues with the portability of the parameters of the model. Currently, new approaches are being developed in order to cope with these difficulties.

Bibliography

[AHSW61] C. B. Anfinsen, E. Haber, M. Sela, and F. H. Jr. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Nat Acad Sci*, 47(9):1309–1314, 1961.

[ANW07] N. Andrusier, R. Nussinov, and H. J. Wolfson. Firedock: fast interaction refinement in molecular docking. *Proteins*, 69(1):139–159, 2007.

[Bal07] R. Baldwin. Energetics of protein folding. *J Mol Biol*, 371:283–301, 2007.

[Bal14] R. L. Baldwin. Dynamic hydration shell restores kauzmann's 1959 explanation of how the hydrophobic factor drives protein folding. *Proc Nat Acad Sci*, 111(36):13052–13056, 2014.

[BB11] T. R. Beattie and S. D. Bell. Molecular machines in archaeal dna replication. *Curr Op in Chem Biol*, 15(5):614–619, 2011.

[BBB+09] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello. Plumed: a portable plugin for free energy calculations with molecular dynamics. *Computer Phys Comm*, 180:1961, 2009.

[BBO+83] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamic calculations. *J Comp Chem*, 4(2):187–217, 1983.

[BBS11] P. Benkert, M. Biasini, and T. Schwede. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27(3):343–350, 2011.

[Bia12] W. Bialek. *Biophysics: searching for principles*. Princeton University Press, 2012.

[BN76] A. Ben-Naim. Statistical potentials extracted from protein structures: are these meaningful potentials? *J Chem Phys*, 107(9):3698–3706, 1976.

[BPvG+84] H. J. C. Berendsen, J. P .M. Postma, W. F. van Gunsteren, A. Di Nola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J Chem Phys*, 81(8):3684–3690, 1984.

[BRW95] N. S. Boutonnet, M. J. Rooman, and S. J. Wodak. Automatic analysis of protein conformational changes by multiple linkage clustering. *J Mol Biol*, 253(4):633–647, 1995.

[BS97] G. P. Brady and K. A. Sharp. Entropy in protein folding and in protein-protein interactions. *Curr Op in Struct Biol*, 7:215–221, 1997.

[BST09] P. Benkert, T. Schwede, and S. C. E. Tosatto. Qmeanclust: Estimation of protein model quality by combining a composite scoring function with structural density information. *BMC Struct Biol*, 9:35, 2009.

[BTS08] P. Benkert, S. C. E. Tosatto, and D. Schomburg. Qmean: a comprehensive scoring function for model quality assessment. *Proteins*, 71(1):261–277, 2008.

[Cat08] A. Caticha. Lectures on probability, entropy and statistical theory. *arXiv, arXiv:0808.0012v1*, 2008.

[CC84] R. Constanciel and R. Contreras. Self-consistent field-theory of solvent effects representation by continuum models - introduction of desolvation contribution. *Theor Chim Acta*, 65:1–11, 1984.

[CGL+12] P. Cossio, D. Granata, A. Laio, F. Seno, and A. Trovato. A simple and efficient statistical potential for scoring ensembles of protein structures. *Sci Rep*, 2, 2012.

[Cha05] D. Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437:640–647, 2005.

[Cho74] C. Chothia. Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248:338–339, 1974.

[CKT09] D. Cozzetto, A. Kryshtafovych, and A. Tramontano. Evaluation of casp8 model quality predictions. *Proteins*, 77:157–166, 2009.

[Clo00] G. M. Clore. Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear overhauser enhancement data and dipolar couplings by rigid body minimization. *Proc Nat Acad Sci*, 97(16):9021–9025, 2000.

[Con83] M. L. Connolly. Analytical molecular surface calculation. *J App Cryst*, 16(5):548–558, 1983.

[Cos11] P. Cossio. *Protein physics by advanced computational techniques: conformational sampling and folded state discrimination*. PhD thesis, SISSA, 2011.

[CPD12] Y. Chebaro, S. Pasquali, and P. Derreumaux. The coarse-grained opep force field for non-amyloid and amyloid proteins. *J Phys Chem*, 116(30):8741–8752, 2012.

[CT15] A. Contini and G. Tiana. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. *J Chem Phys*, 143:025103, 2015.

[DB02] B. N. Dominy and C. L. Brooks. Identifying native-like protein structures using physics-based potentials. *J Comp Chem*, 23:147–160, 2002.

[DBB03] C. Dominguez, R. Boelens, and A. Bonvin. Haddock: a protein-protein docking approach based on biochemical and/or biophysical information. *J Am Chem Soc*, 125(7):1731–1737, 2003.

[DE10] R. Dintyala and R. Elber. Pie - efficient filters and coarse grained potentials for unbound protein-protein docking. *Proteins*, 78(2):400–419, 2010.

[DNRB94] A. Di Nola, D. Roccattano, and H. J. C. Berendsen. Molecular dynamics simulations of the docking of substrates to proteins. *Proteins*, 19(3):174–182, 1994.

[DNW02] D. Duhovny, R. Nussinov, and H. J. Wolfson. Efficient unbound docking of rigid molecules. In Guigó R and Gusfield D, editors, *Proceedings of the 2^{nd} workshop on algorithms in Bioinformatics(WABI)*, volume 2452, pages 185–200. 2002.

[DSZ+12] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *J Phys Chem B*, 116(29):8494–8503, 2012.

[EKS83] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans on Inf Theory*, 29(4):551–559, 1983.

[ESDW+08] U. Emekli, D. Schneidman-Duhovny, H. J. Wolfson, R. Nussinov, and T. Haliloglu. Hingeprot: Automated prediction of hinges in protein structures. *Proteins*, 70(4):1219–1227, 2008.

[FB97] R. Fraczkiewicz and W. Braun. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comp Chem*, 19(3):319–333, 1997.

[FCS+10] S. J. Fleishman, J. E. Corn, E. M. Strauch, T. A. Whitehead, I. Andre, J. Thompson, J. J. Havranek, R. Das, P. Bradley, and D. Baker. Rosetta in capri rounds 13-19. *Proteins*, 78(15):3212–3218, 2010.

[FS02] A. Fernandez and H. A. Scheraga. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc Nat Acad Sci*, 100(1):113–118, 2002.

[FS03] A. Fiser and A. Sali. Modeller: Generation and refinement of homology-based protein structure models. *Methods Enzymol*, 374:461+, 2003.

[GLC94] M. Gerstein, A. M. Lesk, and C. Chothia. Structural mechanisms for domain movements in proteins. *Biochemistry*, 33(22):6739–49, 1994.

[GMW+03] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*, 331(1):281–299, 2003.

[Gol02] E. Golemis. *Protein-protein interactions: a molecular cloning manual*. Cold Spring Harbor Laboratory Press, 2002.

[Gre81] J. Greer. Comparative model-building of the mammalian serine proteases. *J Mol Biol*, 153:1027–1042, 1981.

[GWA01] E. J. Gardiner, P. Willett, and P. J. Artymiuk. Protein docking using a genetic algorithm. *Proteins*, 44(1):44–56, 2001.

[HCK02] C. D. Hu, Y. Chinenov, and T. K. Kerppola. Visualization of interactions among bzip and rel family proteins in living cells using bimolecular fluorescence complementation. *Mol Cell*, 9(4):789–798, 2002.

[HDS96] W. Humphrey, A. Dalke, and K. Schulten. Vmd - visual molecular dynamics. *J Molec Graphics*, 14(1):33–38, 1996.

[HGMO+06] V. J. Hilser, B. E. Garcia-Moreno, T. G. Oas, G. Kapp, and S. T. Whitten. A statistical thermodynamic model of the protein ensemble. *Chem Rev*, 106:1545–1558, 2006.

[HKVdSL08] B. Hess, C. Kutzner, D. Van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Th Comp*, 4(3):435–447, 2008.

[HN95] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268:1144–1149, 1995.

[How00] C. Howson. *Hume's problem - Induction and the justification of belief*. Oxford University Press, 2000.

[HPM+08] H. Hwang, B. Pierce, J. Mintseris, J. Janin, and Z. P. Weng. Protein-protein docking benchmark version 3.0. *Proteins*, 73:705–709, 2008.

[HSL95] E. S. Huang, S. Subbiah, and M. Levitt. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J Mol Biol*, 252:709–720, 1995.

[HVJW10] H. Hwang, T. Vreven, J. Janin, and Z. Weng. Protein-protein docking benchmark version 4.0. *Proteins*, 78:3111–3114, 2010.

[Jan10] J. Janin. Protein-protein docking tested in blind predictions: the capri experiment. *Mol Biosyst*, 6:2351–2362, 2010.

[Jay03] E. T. Jaynes. *Probability theory - The logic of science*. Cambridge University Press, 2003.

[JHM+03] J. Janin, K. Henrick, J. Moult, L. Ten Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak. Capri: a critical assessment of predicted interactions. *Proteins*, 52(1):2–9, 2003.

[JT97] S. Jones and J. M. Thornton. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol*, 272(1):121–132, 1997.

[JTR88] W. L. Jorgensen and J. Tirado-Rives. The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc*, 110(6):1657–1666, 1988.

[JW78] J. Janin and S. Wodak. Conformation of amino acid side-chains in proteins. *J Mol Biol*, 125(3):357–386, 1978.

[Kau59] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv Prot Chem*, 14:1–63, 1959.

[KB04] S. Kojima and D. F. Blair. The bacterial flagellar motor: structure and function of a complex molecular machine. *Int Rev of Cyt*, 233:93–134, 2004.

[KBD+58] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662–666, 1958.

[KKSE+92] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Nat Acad Sci*, 89(6):2195–2199, 1992.

[Kos58] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc Nat Acad Sci*, 44(2):98–104, 1958.

[KRB+02] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Ann Rev Phys Chem*, 53:291–318, 2002.

[KRB+04] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J Comp Chem*, 13(8):1011–1021, 2004.

[KS83] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[KTS+14] M. A. Kasimovaa, M. Tareka, A. K. Shaytanb, K. V. Shaitanb, and L. Delemotte. Voltage-gated ion channel modulation by lipids: Insights from molecular dynamics simulations. *Biochim et Biophys Acta*, 1838(5):1322–1331, 2014.

[KTSW11] K. V. Klenin, F. Tristram, T. Strunk, and W. Wenzel. Derivatives of molecular surface area and volume: simple and exact analytical formulas. *J Comp Chem*, 32(12):2647–2653, 2011.

[LCCJ99] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285:2177–2198, 1999.

[LDA+03] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by cα geometry: ϕ,ψ and cβ deviation. *Proteins*, 50(3):437–450, 2003.

[Lev69] C. Levinthal. How to fold graciously. *Mossbauer spectroscopy in biological systems: proceedings of a meeting held at Allerton House, Monticello, Illinois*, pages 22–24, 1969.

[Lev92] M. Levitt. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol*, 226:507–533, 1992.

[LF00] I. Luque and E. Freire. Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins*, Suppl 4:63–71, 2000.

[LK99] T. Lazaridis and M. Karplus. Effective energy functioin for proteins in solution. *Proteins*, 35:133–152, 1999.

[LLMP+12] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS ONE*, 7(2):e32131, 2012.

[LLPDS11] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334:517–520, 2011.

[LP02] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc Nat Acad Sci*, 99(20):12562–12566, 2002.

[LR71] B. Lee and F. M. Richards. Interpretation of protein structures - estimation of static accessibility. *J Mol Biol*, 55:379, 1971.

[LT13] S. Lui and G. Tiana. The network of stabilizing contacts in proteins studied by coevolutionary data. *J Chem Phys*, 139:155103, 2013.

[MDT09] J. Maupetit, P. Derreumaux, and P. Tuffery. Pep-fold: an online resource for de novo peptide structure prediction. *Nuc Ac Res*, 37:W498–W503, 2009.

[MFK+09] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano. Critical assessment of methods of protein structure prediction, round viii. *Proteins*, 77(S9):1–4, 2009.

[MJ85] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasichemical approximation. *Macromolecules*, 18:534–552, 1985.

[MJ96] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256:623–644, 1996.

[MLLW05] R. Méndez, R. Laplae, M. F. Lensink, and S. J. Wodak. Assessment of capri predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60:150–169, 2005.

[MPL+11] F. Morcos, A. Pagnani, B. Lunta, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, , and M. Weigt. Awsem-md: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *Proc Nat Acad Sci*, 108:E1293, 2011.

[MR05] D. Mustard and D. W. Ritchie. Docking essential dynamics eigenstructures. *Proteins*, 60(2):269–274, 2005.

[MRP+01] J. G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny, and L. F. Ten Eyck. Protein docking using continuum electrostatics and geometric fit. *Protein Eng*, 14(2):105–113, 2001.

[MRY+07] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. The martini forcefield: coarse grained model for biomolecular simulations. *J Phys Chem*, 11(27):7812–7824, 2007.

[NL01] S. Y. Noskov and C. Lim. Free energy decomposition of protein-protein interactions. *Biophys J*, 8(2):737–750, 2001.

[NT03] I. Nooren and J. M. Thornton. Diversity of protein–protein interactions. *The EMBO Journal*, 22(14):3486–3492, 2003.

[OMJ+97] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath - a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[PHW11] B. G. Pierce, Y. Hourai, and Z. Weng. Accelerating protein docking in zdock using an advanced 3d convolution library. *PLoS ONE*, 6(9):e24657, 2011.

[PKWM00] P. N. Palma, L. Krippahl, J. E. Wampler, and J. J. Moura. Bigger: a new (soft) docking algorithm for predicting protein interactions. *Proteins*, 39(4):372–384, 2000.

[PR81] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: a new molecular dynamics method. *J App Phys*, 52:7192–7190, 1981.

[PW07] B. Pierce and Z. Weng. Zrank: reranking protein docking predictions with an optimized energy function. *Proteins*, 67(4):1078–1086, 2007.

[QDWL11] K. Qin, C. Dong, G. Wu, and N. A. Lambert. Inactive-state preassembly of gq-coupled receptors and gq heterotrimers. *Nature Chem Bio*, 7(11):740–747, 2011.

[RF10] D. Rykunov and A. Fiser. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics*, 11, 2010.

[Ric77] F. M. Richards. Areas, volumes, packing and protein structure. *Annu Rev Biophys Bioeng*, 6:151–176, 1977.

[Ric84] T. J. Richmond. Solvent accessible surface area and excluded volume in proteins. analytical equations for overlapping spheres and implications for the hydrophobic effect. *J Mol Biol*, 178:73–89, 1984.

[RSMB04] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods Enzymol*, 383:66–93, 2004.

[RTP+13] V. A. Roberts, E. E. Thompson, M. E. Pique, M. S. Perez, and L. F. Ten Eyck. Dot2: macromolecular docking with improved biophysical models. *J Comp Chem*, 34:1743–1758, 2013.

[SB93] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234:779–815, 1993.

[Sch04] T. A. Schroer. Dynactin. *Annu Rev Cell Dev Biol*, 20:759–779, 2004.

[Sco00] R. Scozzafava. The role of probability in statistical physics. *Transp Th and Stat Phys*, 29:107–123, 2000.

[SGB+14] D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. B. Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. Tak, P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC14)*, pages 41–43, 2014.

[SGS+15] E. Sarti, D. Granata, F. Seno, A. Trovato, and A. Laio. Native fold and docking pose discrimination by the same residue-based scoring function. *Proteins*, 83(4):621–630, 2015.

[Sha49] C. E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[Sip90] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. *J Mol Biol*, 213:859–883, 1990.

[SKHB97] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268:209–225, 1997.

[SO99] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *J Comp Chem*, 314(1-2):141–151, 1999.

[SP00] M. Shirts and V. J. Pande. Screen savers of the world unite! *Science*, 290(5498):1093–1094, 2000.

[SR73] A. Shrake and J. A. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *J Mol Biol*, 79(2):351–371, 1973.

[SR95] R. Srinivasan and G. D. Rose. Linus: a hierarchic procedure to predict the fold of a protein. *Proteins*, 22:81–99, 1995.

[SRK+99] K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34(1):82–95, 1999.

[SUS07] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Mol Sys Biol*, 3(88), 2007.

[SZC+13] E. Sarti, S. Zamuner, P. Cossio, A. Laio, F. Seno, and A. Trovato. Bachscore. a tool for evaluating efficiently and reliably the quality of large sets of protein structures. *Computer Phys Comm*, 184(12):2860–2865, 2013.

[TBB+14] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi. Plumed2: New feathers for an old bird. *Computer Phys Comm*, 185:604, 2014.

[TBM+03] J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, 52(1):76–87, 2003.

[TCMS06] A. Trovato, F. Chiti, A. Maritan, and F. Seno. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput Biol*, 2(12):e170, 2006.

[TD96] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol*, 257:457–469, 1996.

[TS76] S. Tanaka and H. A. Scheraga. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9:945–950, 1976.

[VBW94] A. Varshney, F. P. Jr. Brooks, and W. V. Wright. Computing smooth molecular surfaces. *IEEE Comp Graphics and App*, 14(5):19–25, 1994.

[VHK13] S. Vajda, D. R. Hall, and D. Kozakov. Sampling and scoring: a marriage made in heaven. *Proteins*, 81(11):1874–1884, 2013.

[VHW11] T. Vreven, H. Hwang, and Z. Weng. Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci*, 20(9):1576–1586, 2011.

[VRE13] S. Viswanath, D. V. S. Ravikant, and R. Elber. Improving ranking of models for protein complexes using side chain remodeling and atomic potentials. *Proteins*, 81(4):592–606, 2013.

[WFLS04] K. Wang, B. Fain, M. Levitt, and R. Samudrala. Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Bioinformatics*, 6(8), 2004.

[WN00] H. J. Wolfson and R. Nussinov. Geometrical docking algorithms. a practical approach. In Clifton NJ, editor, *Methods in molecular biology*, volume 143, pages 377–397. 2000.

[WSFB05] C. Wang, O. Schueler-Furman, and D. Baker. Improved side-chain modeling for protein-protein docking. *Protein Sci*, 14(5):1328–1339, 2005.

[WSS99] J. Weiser, P. S. Shenkin, and W. C. Still. Approximate atomic surfaces from linear combinations of pairwise overlaps (lcpo). *J Comp Chem*, 20(2):217–230, 1999.

[WSTT14] I. Walsh, F. Seno, S. C. E. Tosatto, and A. Trovato. Pasta 2.0: an improved server for protein aggregation prediction. *Nuc Ac Res*, 42:W301–W307, 2014.

[Wut01] K. Wuthrich. The way to nmr structures of proteins. *Nature*, 8(11):923–925, 2001.

[XLX04] G. B. Xu, M. Li, and Y. Xu. Protein threading by linear programming: theoretical analysis and computational results. *J of Comb Opt*, 8(4):403–418, 2004.

[XTN97] D. Xu, C. J. Tsai, and R. Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng*, 10(9):999–1012, 1997.

[YD96] K. Yue and K. A. Dill. Folding proteins with a simple energy function and extensive conformational searching. *Protein Sci*, 5:254–261, 1996.

[YJEP07] C. J. Yu, T. Joachims, R. Elber, and J. Pillardy. Support vector training of protein alignment models. In Speed T and Huang H, editors, *RECOMB 2007*, pages 253–267. 2007.

[YPVV08] S. Yang, I. C. Paschalidis, P. Vakili, and S. Vajda. Protein docking by the underestimation of free energy funnels in the space of encounter complexes. *PLoS Comput Biol*, 4(10):e1000191, 2008.

[Zam15] S. Zamuner. *Local sampling and statistical potentials for scoring protein structures*. PhD thesis, Universitá degli Studi di Padova, 2015.

[Zha08] Y. Zhang. I-tasser server for protein 3d structure prediction. *BMC Bioinformatics*, 9(40), 2008.

[ZS01] H. X. Zhou and Y. B. Shan. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, 44(3):336–343, 2001.