Università degli Studi di Milano – Bicocca

Dottorato di Ricerca in Scienze Chimiche – XXIII Ciclo

Predicting the binding modes of protein complexes:

NEW STRATEGIES FOR MOLECULAR DOCKING

Annalisa Bordogna

Tutor: Prof. Laura Bonati

Coordinatore: Prof. Franca Morazzoni

Anno Accademico 2009/2010

Cover design: Annalisa Bordogna & Luca Masotto

“Nobody believes in simulation models except their developers…

Everybody believes in experimental data except who collected them.”

(*Gaylon S. Campbell*)

Ai miei genitori

A Luca

Table of contents

List of abbreviations

3D	three-dimensional
CAPRI	**C**ritical **A**ssessment of **PR**edicted **I**nteractions
CASP	**C**ritical **A**ssessment of techniques for protein **S**tructure **P**rediction
CATH	**C**lass, **A**rchitecture, **T**opology, **H**omologous Superfamily classification
dRMSD	**d**istance **R**oot **M**ean **S**quare **D**eviation
EM	**E**nergy **M**inimization
FFT	**F**ast **F**ourier **T**ransform
MC	**M**onte **C**arlo
MD	**M**olecular **D**ynamics
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
NMR	**N**uclear **M**agnetic **R**esonance
PDB	**P**rotein **D**ata **B**ank
RMSD	**R**oot **M**ean **S**quare **D**eviation

Chapter 1

Introduction

1.1 General introduction

Proteins are at the basis of life. They are essential parts of every organism, are present in every cell and cellular compartment, and are often protagonists in crucial biological processes. They have structural, mechanical and transport functions, play an important role in cell signalling, are involved in immune responses and, as enzymes, they catalyze biochemical reactions. Moreover, proteins work cooperatively. Their association with other proteins or peptides and their interaction with small molecules, lipids or nucleic acids are fundamental for their function. Therefore, understanding the interaction network of each protein is fundamental for the comprehension of its functions and of the biological mechanisms in which it is involved. This can also provide crucial information for the elucidation of several disease processes, allowing the design of new drugs aimed at inhibiting disease-generated interplays or at restoring interactions that are vital for the organism.

As proteins functions are determined by their three-dimensional structure, the availability of high-resolution 3D structures of protein complexes are essential for a complete understanding of the biochemical nature of the process in which they are involved, and to facilitate the design of compounds that might influence it.

During the last decades, two experimental techniques, X-ray diffraction and nuclear magnetic resonance (NMR), have been widely used and much improved to determine 3D structures of proteins, both as monomers and in complex with small molecules and ions, other proteins and peptides, or DNA and RNA. Their increasing success in resolving protein structures is highlighted by the increasing number of depositions in the Protein Data Bank (PDB) [1] every year.

However, despite the large improvements and the excellent results achieved by these techniques (for example, the X-ray determination of the ribosome structure [2]), they still have some limits. In fact, the crystallization of the protein (or complex) for X-ray experiments is sometimes very difficult, since the growth of a crystal is strongly dependent not only to the temperature, pH and solvent conditions, but also to the nature of the protein and to the absence/presence of ions and ligands. NMR, on the other hand, is suitable only for determining the structures of low molecular weight systems [3], thus it is mostly used for resolving monomers.

Moreover, the nature of the interactions within protein complexes can affect the successful application of these experimental techniques. For example, the majority of the associations that take place in the cell between proteins, or between proteins and other molecules, are transient. The transient nature of these associations is essential for life, since it allows the organism to answer in a fast and reversible way to environmental or metabolic changes. Unfortunately, their nature makes also these interactions very likely to disrupt under experimental conditions, thus, even if they are the most interesting to understand, they are also the most difficult to study.

Due to the experimental restrictions described above and although a number of structural genomics projects [4, 5] have been launched over the years, the number of protein structures deposited in the PDB is still limited (around 70,000 at the end of December 2010, as reported in the PDB). On the contrary, the number of known protein sequences is much higher (around 523,000 at the end of December 2010, as reported in the UniProtKB/Swiss-Prot database [6, 7]), thanks to several genomic sequencing projects [8]. Consequently, in many cases it is necessary to predict the 3D structure of proteins whose sequence is known by computational approaches. To this aim, in the last years several methods aimed at predicting the 3D structures of both protein monomers and proteins interacting with other molecules were developed.

1.2 Computational methods to predict protein structures

The most used bioinformatics methods that allow the prediction of protein monomers structures are: protein folding algorithms, fold recognition methods, fragment-based modelling and homology modelling.

Folding algorithms (called also ‘*ab initio*’ methods) exploit the observation that the same protein sequence adopts almost invariably the same fold [9]. This suggests that such conformation corresponds to the global minimum of free energy under physiological conditions. Thus, these methods perform an extensive sampling of the conformational space of the protein, for example by Molecular Dynamics, in order to find the energy minimum associated to the native protein structure.

Fold recognition (or 'threading'), on the other hand, takes advantage of the observation that the majority of the proteins fold into a limited number of topologies [10, 11]. These methods 'thread' the sequence of the protein whose structure has to be predicted onto a large number of different experimental structures and select the best sequence-structure match [12, 13].

Fragment-based algorithms are based on the structure similarity at a local level (some residues) between proteins that are apparently not related [14, 15]. On this basis, they function assembling fragments taken from experimental structures and then refining the structure obtained.

Homology modelling (or 'comparative modelling') [16] derives from the observation that evolutionary related (homologous) proteins have similar structures, even if they have quite different sequences [17]. The structure of homologous proteins are therefore used as a template to model the structure required.

This last method is one of the most widely used in protein structure prediction. This is because of the great number of protein sequences currently available in public databases and because it has been demonstrated that the expected reliability of the final model can be estimated *a priori* and depends on the distance in homology between the protein whose structure has to be predicted (target) and the one on which the structure is modelled (template) [17].

The classical procedure for building an homology model can be summarized as follows:

1. identify a suitable template among proteins of known structure that are homologous to the target;
2. build a sequence alignment between target and template;
3. assign the atomic coordinates of the backbone and of the identical side-chains of the template to the target, according to the sequence alignment;
4. model the remaining side-chains and the non-aligned backbone regions;
5. refine the structure obtained.

Among these stages, the first two steps are the most crucial for a high-quality final model. In particular, the selected template should be a protein whose experimental structure is available and which is the best evolutionary related to the target. However, once there is a choice of different possible templates with similar sequence identities for a particular target, it may be difficult to reliably choose the most similar [18, 19]. The sequence alignment, on the other hand, has to be biologically meaningful and should detect the correspondence between amino acids that matches the evolutionary history of the protein family [20, 21].

Moreover, the use of different modelling strategies and final refinements, starting from the same template and alignment, can result in models at different quality. In fact, a difference in modelling loops and side-chains frequently leads to models with the same overall quality, but different local accuracy. This difference may influence the results of subsequent studies on the models, for example docking calculations, that require a very accurate description of the binding site in order to obtain reliable results. Consequently, critical testing of the quality of protein models is essential if they are to be applied with confidence.

To assess the global and local quality of protein structural models, a number of quality indices have been developed over the years [22, 23]. These indices are usually based on the results of statistical analyses on the whole PDB (or on some subsets of it), taking into account a number of structural features of experimental structures, and they assess if the characteristics of the structures to analyze fall in the average of the reference structures. Another class of model quality indices is composed by methods that generate statistical potentials from a set of PDB structures and use them to evaluate the structure that has to be examined.

A benchmarking of methods for protein structure prediction is carried on every two years, through a world-wide experiment named CASP (Critical Assessment of techniques for protein Structure Prediction) [24, 25], aimed at establishing the current state of the art, identifying what progress has been made over the years and highlighting where future efforts may be most productively focused. This blind modelling experiment has the great merits of having raised the issue of objective evaluation of structure prediction methods, prompting the development of model quality indices, and of having encouraged the ideation of new and better performing modelling approaches.

1.3 Molecular docking calculations

Molecular docking is a computational approach that allows the prediction of the structure of a receptor-ligand complex, starting from the structures of the two interacting partners [26]. The biggest of the two interacting molecules is classified as ‘receptor’, while the smallest is the ‘ligand’. In principle, these two molecules can be proteins, small molecules, RNA, DNA, ect. However, at the present time, the most studied interactions are those that occur among proteins (protein-protein docking) and between proteins and small molecules (ligand-protein docking).

The majority of the currently available docking methods work in two steps:

1. sampling: search in the conformational space of the relative orientations of the interacting partners, in order to generate a set of tentative solutions (‘docking poses’);
2. refinement and scoring of the generates poses, through the use of a force field adequate to select the pose that is the most similar to the native structure of the complex.

In principle, the sampling should take into account both the six degrees of translational and rotational freedom of one body relative to another (intermolecular conformational space) and the conformational degrees of freedom of the ligand and of the receptor as single molecules (intramolecular conformational space). As for ranking the poses, the scoring function should predict the loss of free energy occurring upon binding, accurately calculating both the interaction and solvation energy of the complex and of the partners considered as single and separate molecules. With respect to the achievement of these goals, the two most popular classes of docking approaches, ligand-protein and protein-protein, although developed roughly in the same years, have currently different limitations and can achieve different degrees of accuracy.

1.3.1 Ligand-protein docking

Ligand-protein docking is extensively used by pharmaceutical companies. This is why, over the years, enormous efforts have been made to improve the corresponding software. In fact, these calculations help in cutting the time and financial investment required for the development of novel pharmaceutical agents. In particular, these methods have proven to be useful at different stages of the drug discovery process for three main purposes [27]:

- predicting the binding mode of a known active ligand;
- predicting the binding affinities of related compounds from a known active series;
- identifying new ligands using virtual screening.

The successful prediction of a ligand binding mode is the field where the most success has been achieved. While the first docking algorithms treated both ligand and protein as rigid bodies, a first great improvement has been the introduction of sampling algorithms able to perform a near-complete search in the intramolecular conformational space of the ligand [28]. Today, most docking programs treat the ligand as flexible with a rigid (or nearly rigid) receptor structure, while receptor flexibility remains one of the major challenges in this field [29, 30]. A number of reviews report the performance of different docking methods in terms of ability of reproducing the native binding mode [31-35] and show that the best programs do predict the experimental pose in most of the times.

On the other hand, scoring functions, that can be both physical- or knowledge-based, have still to be improved. In fact, they are frequently unable to distinguish near-native poses from the others [35, 36]. Moreover, they usually strongly depend on the physico-chemical features of the set of complexes they were trained on [31-35]. The limits of the scoring functions determine also the limits of the currently available docking programs to reliably reproduce even the rank of binding affinities, when considering a set of compounds docked into a single protein [27, 34]. In spite of these limitations, a number of successes are reported in literature about the application of ligand-protein docking to the problem of identifying new ligands through virtual screening (see 37 for a review). Moreover, the enhancements of computer performance allow now the screening of hundreds of thousands molecules in a reasonable time and several docking algorithms have been optimized and have been automated specifically for this use, which makes ligand-protein docking calculations a sort of routinary process in drug discovery.

1.3.2 Protein-protein docking

The most substantial improvements in protein-protein docking algorithms, on the other hand, have been prompted by a world-wide blind docking experiment, called CAPRI (Critical Assessment of PRedicted Interactions) [38, 39]. This competition, similarly to CASP, aims at periodically assessing the state of the art in protein-protein (and protein-nucleic acids) docking and at underlining which progresses are made, trying to understand the avenues of improvement that the docking community should open. Each CAPRI experiment (round) is composed by two parts, which are designed to separately assess the sampling and scoring stage of the currently available protein-protein docking methods.

As CAPRI results point out, the sampling step of protein-protein docking algorithms is not always successful, especially when conformational changes occur within the proteins upon binding [40, 41]. In fact, due to the complexity of the search into the inter- and intramolecular conformational space and to the huge computational time requested for such a sampling, each protein-protein docking method suffers from some kind of approximation. Some algorithms (Fast Fourier Transform approaches, FFT, also called '*ab initio*') treat the proteins as rigid bodies and this allows them to effectively sample the intermolecular space of the complex. Some other approaches (that make use of Monte Carlo search, Molecular Dynamics or Energy Minimization), instead, treat the proteins as flexible, but limit the intermolecular space search only to the areas of the protein surface that may be interacting patches. These should be predicted by bioinformatics methods or should be derived from experimental data [42, 43]. Recently, it has been developed the idea that the limitations of the two sampling approximations would be overcome by the combined use of docking methods belonging to these two different philosophies [44].

On the other hand, scoring functions are currently neither useful for predicting binding affinities, as recently demonstrated on a benchmark of protein-protein complexes [45], nor, at the present state of the art, they can discriminate between proteins that can or cannot bind to another protein, as assessed in a very recent CAPRI round (confidential data).

1.4 The use of homology models in docking calculations

The recent improvements in homology modelling techniques [46, 47], the development of automated methods and the availability of models repositories like for example SWISS-MODEL [48], the Protein Model Portal [49] and Modbase [50], have greatly extended the use of homology models for subsequent experiments. In particular, due to the lack of experimental structures for a number of pharmaceutical target proteins, the use of protein models in docking, particularly for drug design applications, is rapidly growing [51]. As a consequence, a topic of great interest is the assessment of the potentialities and limits of the use of protein models in ligand-protein docking.

Several recent studies already dealt with the problem of identifying the relationship between model quality and ligand-protein docking results accuracy, in some specific cases of high throughput screening [52-58]. In those studies, the accuracy of the models was indirectly estimated by the sequence identity with the template, whereas the docking performance was quantified in terms of enrichment of known active compounds against a background of molecules. The subject of those analysis were proteins that are pharmaceutically relevant targets and their aim was to assess the reliability of high throughput screening when a model of the target is used. From those studies, a wide range of trends was obtained and a clear relationship between sequence identity and enrichment factor was not observed. However, from some of those works, a generally accepted 'rule' emerged: models built with over 50% sequence identity with the template are accurate enough for safe docking studies [51, 54, 59], since they are usually able to globally reproduce the reference structure with high accuracy.

Although the overall model quality is indeed related to the expected structural divergence between template and target [17], the above rule is not always valid when applied to docking experiments, because additional factors play a role in determining the actual accuracy of a modelled structure, as already mentioned in Paragraph 1.2. Moreover, several studies were based on the hypothesis that the global Root Mean Square Deviation (RMSD) of the model from the experimental structure can be directly related to the ability of docking methods to reproduce the ligand pose in a specific binding site. This fact, though, has never been rigorously demonstrated: a globally correct model can indeed include a bad description of the active site and *vice versa*. This would lead to accurate docking predictions for globally not accurate models with a modelled binding site that is adherent to the native and, on the other way round, to wrong predictions for globally very accurate models with a binding site geometry very different from the native.

Only a large-scale benchmarking study recently published [57] suggested some techniques to exploit comparative models at best in molecular docking screenings. However, none of the model quality indices tested in that study (among them, the identity percentage for the whole sequence and for the binding site) appeared suitable to predict the accuracy of ligand docking with acceptable reliability.

From these observations it clearly emerges that deriving general relationships between model quality and docking accuracy is central to a more effective use of docking simulations. In the most desirable scenario docking accuracy would be predicted directly from the quality of the protein model assessed by employing the model quality indices already available to the modelling community.

It is conceivable that protein models will be exploited in future years also in protein-protein docking studies, as suggested by the recent CAPRI trend of assessing docking predictions on protein models [60, 61]. In this field, no extensive study has been made in this direction until now. Therefore, it is now of great interest to assess both to which extent models can be used in protein-protein docking to safely obtain accurate results – for example defining a target-template sequence identity percentage above which accurate docking results are surely obtained – and to investigate if it is possible to predict the accuracy of docking results on the basis of the quality of the modelled structure, indicated by standard model quality indices.

1.5 Objectives of the thesis

This thesis was mainly focused on overcoming the limitations of the sampling stage of ligand- and protein-protein docking methods and exploring the potentialities of combining different computational techniques to broaden the possibility to predict the structure of protein complexes.

For this reason, in the protein-protein docking field, to overcome the still existing limitations of the available algorithms in the sampling stage, I decided to deal with the development of a new strategy for the initial search stage. This was made following a very recent tendency [44, 62-65], that has already proven its utility in the field of protein-protein docking: the combination of docking programs that belong to the two principal protein-protein search approximations, FFT and data-driven docking. This approach allows to sum the strengths of the two search philosophies (fast and almost complete intermolecular sampling for the first, accurate treatment of molecular flexibility for the latter) and limit their weaknesses (lack of consideration of the intramolecular degrees of freedom for the first, dependence on external data and on their accuracy for the latter).

On the other hand, to exploit the potentiality of the recent improvements of the homology modelling techniques [46, 47] both in ligand- and in protein-protein docking, I decided to focus part of my work on the effects of the use of homology models in docking calculations, assessing the relationships between docking results accuracy and model quality (evaluated with the most widely used model quality indices) and developing a strategy to predict the goodness of docking results, on the basis of model quality [66].

1.6 Outline of the thesis

In this introduction, the main motivations for the work presented in the thesis were explained.

The methods and programs used are presented in Chapter 2. In particular, Paragraph 2.1 treats the topic of homology modelling, Paragraph 2.2 presents an overview of model quality indices, and Paragraph 2.3 and Paragraph 2.4 illustrate ligand- and protein-protein docking, respectively.

Chapter 3 is focused on overcoming the current sampling limitations of protein-protein docking methods: it describes ZADDOCK, a new protein-protein docking approach that combines FFT and data-driven search into a whole pipeline.

Chapter 4 and 5 focus on the use of homology models in ligand- and protein-protein docking, respectively. In particular, in both chapters an analysis of the relationships between model quality and docking results accuracy is presented. In Paragraph 4.2 is also reported the description of a strategy to predict the goodness of ligand-protein docking results on the basis of the quality of the model used.

In Chapter 6 the conclusions about this work are drawn and future perspectives in the homology modelling and docking fields are presented.

Chapter 2

Computational methods

In this thesis, a number of programs was used. They were employed for: protein structure prediction; ligand- and protein-protein docking; for evaluating the quality of the models and the accuracy of docking results.

2.1 Protein structure prediction

The structure prediction part of this thesis was performed with the help of two methods: MODELLER [67-69] and I-TASSER [70, 71]. The first is a homology modelling program, the latter is a threading approach.

2.1.1 MODELLER

MODELLER [67-69] uses an automated approach to model protein structures by homology modelling, through the satisfaction of spatial restraints. This program implements different types of approaches for modelling and refinement; in the following paragraphs the basic workflow will be described.

MODELLER works in three stages:

1. alignment of the target sequence to the template structure and generation of the starting model;
2. extraction of spatial restraints;
3. optimization of the model by satisfying spatial restraints.

Alignment of the target sequence to the template structure and generation of the starting model

The sequence of the target is first aligned to the structure of the template(s), following the sequence alignment given in input to the program. An initial model is generated, on the basis of the alignment.

Definition of spatial restraints

The restraints for modelling are derived directly from the alignment between the target sequence and the template structure(s) and mainly involve structural features like for example main-chain and side-chain geometrical properties, residue accessibility and neighbourhood relationships, as well as sequence alignment properties. The complete list is reported in Table 2.1. A statistical analysis of the relationships between pairs of homologous structures is at the basis of the definition of the restraints; this analysis relies on a database of 105 family alignments that included 416 proteins with known 3D structure. The obtained relationships are expressed as conditional probability density functions (pdfs) and are used directly as spatial restraints. The spatial restraints and CHARMM energy terms enforcing proper stereochemistry [72] are combined into an objective function.

Optimization of the model by satisfying the spatial restraints

The final model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function approach employing methods of conjugate gradient and molecular dynamics with simulated annealing.

Table 2.1 – List of MODELLER spatial restraints.

Restraint name	**Variable**
Amino acid residue type	r
Main-chain dihedral angle Φ	Φ
Main-chain dihedral angle Ψ	Ψ
Secondary structure class (residue)	t
Main-chain conformation class (residue)	M
Fractional content of residues in the main-chain conformation class A	α
Side-chain dihedral angle χi , i=1,2,3,4	χi
Side-chain dihedral angle χi class, i=1,2,3,4	ci
Residue solvent accessibility	a
Average accessibility of two residues in one protein	ā
Residue neighbourhood difference between two proteins	s
Average residue neighbourhood difference between two proteins	s̄
Fractional sequence identity between two proteins	i
Cα-Cα distance	d
Difference between two Cα-Cα distances in two proteins	Δd
Main-chain N-O distance	h
Difference between two N-O distances in two proteins	Δh
Average residue Biso	b
Resolution of X-ray analysis	R
Distance of a residue from a gap in alignment	g
Average distance of a residue from a gap	ḡ

Starting from an alignment, MODELLER generates a user-defined number of models, all slightly different one from the other, obtained by varying the initial model.

The selection of the model (or models) to use for subsequent analyses is up to the user. This can be made either by evaluating the models with standard structure quality indices (see Paragraph 2.2) or by considering the values of the MODELLER objective function and selecting the model(s) with the optimal values.

2.1.2 I-TASSER

I-TASSER [70, 71] is a hierarchical protein structure modelling approach which is based on secondary structure-enhanced profile-profile threading alignment (PPA) and employs the iterative implementation of the threading assembly refinement (TASSER) program [73]. The overall procedure is described in Figure 2.1 and is composed of 3 steps:

1. threading of the query sequence through the PDB;
2. first structure assembly, refinement and clustering;
3. structure re-assembly and model selection.

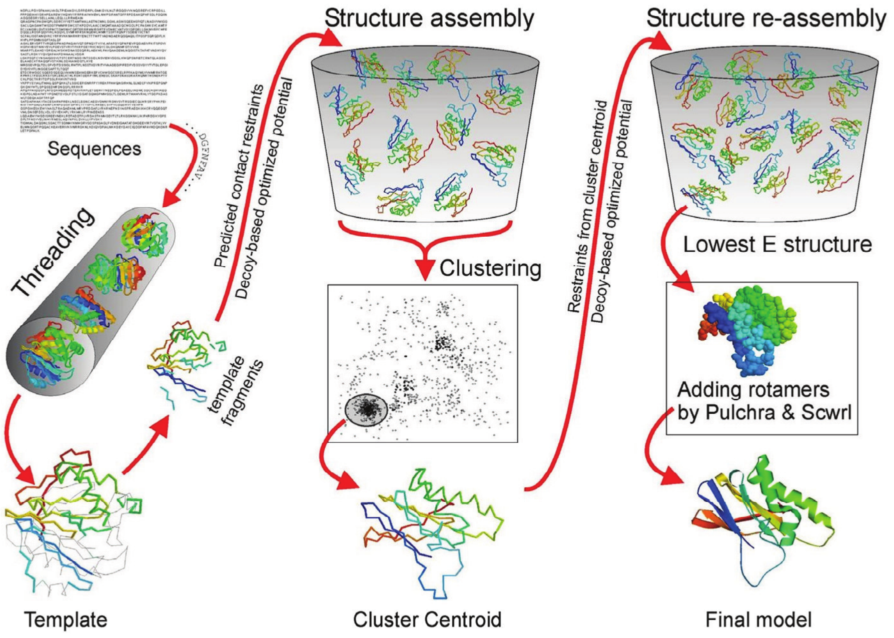

Sequences
Threading
Template
template fragments
Predicted contact restraints
Decoy-based optimized potential
Structure assembly
Clustering
Cluster Centroid
Restraints from cluster centroid
Decoy-based optimized potential
Structure re-assembly
Lowest E structure
Adding rotamers by Pulchra & Scwrl
Final model

Figure 2.1 – The I-TASSER workflow (extracted from the original paper by Wu et al. [70]).

Threading of the query sequence through the PDB

The query sequence is first threaded through the PDB to identify appropriate local fragments, which are then adopted for the structural assembly. The threading method is a profile-profile alignment (PPA) approach, in which an alignment score between each residue of the query sequence and the corresponding residue of the template structure is calculated both on evolutionary and on structural similarity bases. In fact, it takes into account both the probability of substitution between the residues and the match between the secondary structure assignment by DSSP [74] for the template residue and the secondary structure prediction by PSIPRED [75] for the corresponding query residue.

The Needleman-Wunsch dynamic programming algorithm [76] is used to find the best match between the query and template sequences. A position-dependent gap penalty is used: no gap is allowed inside the secondary structure regions, while gap opening (value equal to 7.0) and gap extension (equal to 0.5) penalties apply to other regions and the ending gap penalty is ignored.

First structure assembly, refinement and clustering

In the I-TASSER modelling step, a protein is represented by its Cα atoms and side chain centres of mass (SG). Based on the PPA alignment, the chain is divided into continuous aligned regions (more than five residues), whose local conformation remains unchanged during assembly, and gapped *ab initio* regions. For computational efficiency, the Cα of these *ab initio* residues lie on an underlying cubic lattice; whereas, for maximum accuracy, the Cα of aligned residues are excised from the threading template and are off-lattice. SGs are always off-lattice.

For a given alignment, an initial full-length model is built by connecting the continuous secondary structure fragments through a random walk of Cα-Cα bond vectors of variable lengths. During the initial model-building procedure, only excluded volume and geometric constraints of virtual Cα-Cα bond angles are considered. The side-chain centre of mass is determined by a two-rotamer approximation that depends on whether the local backbone configuration is extended or compact.

The initial full-length models are submitted to parallel-exchange Monte Carlo sampling [77] for assembly/refinement. Two kinds of conformational updates (off-lattice and on-lattice) are implemented. Overall, the tertiary topology varies by the rearrangement of the continuously aligned substructures, where the local conformation of the off-lattice substructures remains unchanged during assembly.

The assembly force field includes predicted secondary structure propensities from PSIPRED, backbone hydrogen bonds, and a variety of statistical short-range and long-range correlations (e. g. consensus predicted side-chain contacts or hydrophobic interactions.) [73]. Moreover, it incorporates a term of predicted accessible surface area (ASA), calculated through a two-state (exposed/bury) neural network (NN) [78], trained on protein structures at high resolution.

Structure re-assembly and model selection

The structure trajectories of the first-round TASSER simulations are clustered by SPICKER [79], an iterative structural clustering program. The cluster centroids are obtained by averaging all the clustered structures after superposition. Following the clustering, the TASSER Monte Carlo simulation is performed again, starting from the cluster centroid conformations. The distance and contact restraints in the second-round TASSER are taken from the combination of the centroid structures and the PDB structures searched by the structure alignment program TM-align [80] based on the cluster centroids. The conformation with the lowest energy in the second round is selected. Finally, the program Pulchra [81] is used to add backbone atoms (N, C, O) and the program Scwrl_3.0 [82] is used to build side-chain rotamers. The sidechain-building procedure by Pulchra and Scwrl does not modify the Cα coordinates.

2.2 Model quality indices

The quality of a protein structural model can be measured by the degree of structural similarity to the native structure, but this implies the availability of the 'answer' to the modelling problem. When this is not the case, the quality is estimated by comparison to the template structure and by conformity to average properties of known protein structures. Recently, moreover, new approaches based on machine learning have been proposed to predict model quality.

In this thesis, models were assessed both by direct comparison to the known native structures ('calculated' indices) and by using indices for model quality estimation and prediction ('predicted' indices).

2.2.1 'Calculated' indices

Direct comparison to the native structure was obtained by structural alignment using three programs: DALILite [83], LGA [84] and ProFit [85]. The former generates a global alignment by optimization of both the number of structurally aligned residues and the overall root mean square deviation (RMSD) of their corresponding Cα atoms. LGA generates many different local superimpositions to detect regions where proteins are similar, providing a combination of two sets of scores for Local/Global Alignment. ProFit exploits an iterative procedure based on the conjugate gradient minimization method that uses a succession of rotations to perform the superimposition of protein structures [86].

According to the structural alignment results, the quality of the models was measured by several indices aimed at evaluating both the global and local accuracy of the structures.

The global quality indices are:

- **RMSD**: the root mean square deviation on Cα atoms included in DALILite structural alignment;
- **DALI_ZSCORE**: the statistical significance of the DALILite alignment compared to a set of unrelated proteins;
- **DALI_ZRATIO**: the DALI_Zscore normalized by the Z-score of structural alignment between model and template. This measure reflects the degree of modelling success as the efficiency to reproduce the native better than the original template [21];
- **LGA_RMSD**: RMSD of the residues of the model superimposed to the corresponding residues of the native structure under the distance cut-off of 4 Å by LGA;
- **GDT_TS** and **GDT_HA**: average of the percentage of residues of the model that can be superimposed to the native structure under a certain distance cut-off. In particular: GDT_TS = (P1+P2+P4+P8); GDT_HA = (P0.5+P1+P2+P4)
  where Pd is the percentage of residues superimposed under the cut-offs d (d = 1, 2, 4, 8 Å for GDT_TS; d = 0.5, 1, 2, 4 Å for GDT_HA);
- **LGA_S**: combined (local and global alignment) LGA score [84] for the LGA alignment.

The local quality was investigated by comparing the geometry of the modelled and native binding sites (for ligand-protein docking cases) or interfaces (for protein-protein docking).

For ligand-protein docking, the residues in the active site were defined as the ones with at least one heavy atom within 4 Å of any of the ligand atoms.

The local quality indices considered for the ligand-protein docking part of this thesis are:

- **RMSD-s**: the root mean square deviation on Cα atoms of the binding site included in DALILite structural alignment;
- **dRMSD-s**: in general, the dRMSD is defined as the root mean square deviation between the corresponding atomic distances in two structures:

$$dRMSD = \sqrt{\frac{\sum_i \sum_j (d_{ij}^a - d_{ij}^b)^2}{N}}$$

(where a and b are the labels of the two structures, d the distance vectors, i and j the indices of the atoms and N is the number of comparisons performed). For the dRMSD-s the sum is performed over the heavy atoms of the side-chains in the active site. Therefore this index records the deviation in the relative positions of the side-chains and it is a suitable complement to the RMSD-s that conversely describes the absolute deviation from the crystallographic geometry;

- **ACS**: the Atomic Contact Score, which evaluates the number of reproduced contacts between the heavy atoms of protein and ligand, after a local superimposition of the model onto the target structure [87]. The ACS evaluates the fraction of the correctly modelled atomic contacts and penalizes the clashes between the modelled binding site atoms and the ligand atoms:

$$ACS = \frac{\sum_{i,j} (Cont_{i,j}^{Xray} - Cont_{i,j}^{m}) - \sum_{i,j} Clash_{i,j}^{m}}{\sum_{i,j} Cont_{i,j}^{Xray}}$$

$$Cont_{i,j} = \begin{cases} 1 & 2.0 \le r_{i,j} \le 4.0 \\ 0 & otherwise \end{cases}$$

$$Clash_{i,j} = \begin{cases} 1 & r_{i,j} \le 1.5 \\ 0 & otherwise \end{cases}$$

where Cont is the number of contacts and Clash of clashes in the crystallographic (Xray) and model (m) structures. The units are in Å.

For the protein-protein docking, the interface backbone was defined as the backbone atoms of the residues within 10 Å of any heavy atom of the interacting partner and the interface side-chains as the side-chain heavy atoms of the residues within 5 Å of any heavy atom of the interacting partner. The local quality indices calculated in this thesis are:

- **iRMSD_bb**: the root mean square deviation on the backbone heavy atoms of the interface calculated from ProFit alignment;
- **iRMSD_sc**: the root mean square deviation on the side-chain heavy atoms of the interface calculated from ProFit alignment;
- **fnat**: fraction of native contacts with the interacting protein reproduced when superimposing the interface of the model onto the interface of the native structure in complex; it was calculated at a distance of 5 Å.

2.2.2 'Predicted' indices

Since the choice of the template greatly affects the modelling results [18], indices that evaluate the model-template similarities (for both structure and sequence, both global and local) were considered in the class of 'predicted' indices. Complementary to these indices is the set of indices either derived by geometrical analysis of the models and comparison with the average of a set of reference structures or based on statistical force fields generated from analyses of high-resolution PDB structures.

The indices employed to evaluate the models by comparison with the templates are:

Global indices:

- **Seq_Id**: the overall percentage of sequence identity between target and template, evaluated on the sequence alignments used for modelling;
- **Seq_Sim**: the overall percentage of sequence similarity between target and template, evaluated on the sequence alignments used for modelling, by using the BLOSUM62 matrix [88];
- **RMSD(t)**: the root mean square deviation on Cα atoms included in the DALILite model-template structural alignment;
- **LGA_RMSD(t)**: RMSD on Cα atoms of the residues of the model superimposed to the corresponding residues of the template structure under the distance cut-off of 4 Å by LGA;
- **LGA_S(t)**: combined (local and global alignment) LGA score on Cα atoms for the LGA model-template alignment.

Local indices for the ligand-protein binding site:

- **Seq_Id-s**: the percentage of sequence identity between target and template in the binding site, evaluated on the sequence alignments used for modelling;
- **Seq_Sim-s**: the percentage of sequence similarity between target and template in the binding site, evaluated on the sequence alignments used for modelling, by using the BLOSUM62 matrix;
- **RMSD-s(t)**: the root mean square deviation on Cα atoms of the binding site included in the DALILite model-template structural alignment.

Local indices for the protein-protein interface:

- **Seq_Id-i**: the percentage of sequence identity between target and template at the interface, evaluated on the sequence alignments used for modelling;
- **iRMSD_bb(t)**: the root mean square deviation on the backbone heavy atoms of the interface calculated from the DALILite model-template alignment.

The 'statistical' predicted quality indices were calculated by using a number of web-servers or programs for the assessment of model (or experimental structure) overall quality:

From the PSVS web-server [89]:

- **MolProbity**: MolProbity clashscore [90], which evaluates the number of overlaps per thousand atoms;
- **Procheck**: Procheck phi-psi score, which checks the stereochemical quality of the phi and psi dihedral angles [91];
- **ProsaII**: ProsaII score, which is based on knowledge-based potentials and evaluates the structures in terms of atom-pair and protein-solvent interactions [92];
- **Verify3D**: Verify3D score, which analyzes the compatibility of a 3D-structure with its own amino acid sequence [93].

From ProQ [94], which is a neural network-based model quality assessment program:

- **LGscore**: predicted score of structural similarity between model and target [95];
- **MaxSub**: predicted score that takes into account the largest number of residues for which all distances between the model and the native structure are shorter than 3.5 Å [96].

From the web-server ModFOLD [97]:

- **MQscore**: it combines scores obtained from different methods using a neural network trained with the TM-score [98].

From the ModEval model evaluation server [99]:

- **DOPE**: Z-score of the DOPE potential, an atomic distance-dependent statistical potential [100];
- **TSVMod_RMSD**: predicted RMSD (all Cα atoms) between model and native structure; it is obtained from the TSVMod method [101], which uses a support vector machine to predict the structural similarity between a model and the corresponding native structure;
- **TSVMod_Over**: predicted fraction of Cα atoms within 3.5 Å of their correct positions in the native structure, also obtained by the TSVMod method [101].

From the QMEAN server for model quality estimation [102]:

- **Qmean**: score derived from statistical potentials taking into account torsion angles, solvation, contacts and terms of agreement between predicted and calculated secondary structure and solvent accessibility [103, 104];
- **Qmean_Z**: Z-score derived by the relation between the Qmean score value for the query model and the scores of a non-redundant set of high-resolution X-rays structures of similar size [105].

2.3 Ligand-protein docking with AutoDock

The ligand-protein docking calculations, in this thesis, were performed by using AutoDock [29, 106, 107]. This program, as many other docking methods, works in 2 stages:

1. sampling of the conformational space of the complex;
2. scoring and ranking of the poses.

2.3.1 AutoDock sampling stage

In this thesis, the Lamarckian genetic algorithms sampling implemented by AutoDock was used. These are genetic algorithms followed by a local search in the conformational space of the ligand, performed on a fraction of the poses generated (a schematic workflow is reported in Figure 2.2).

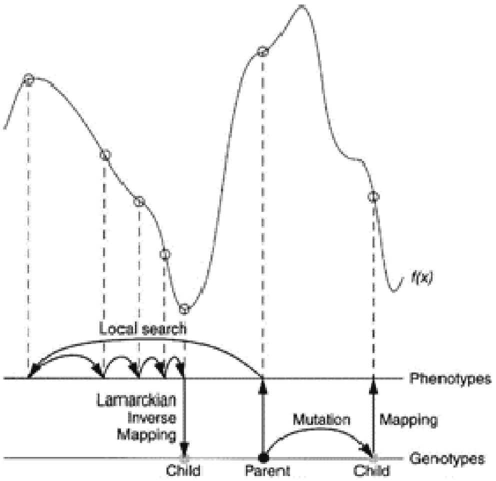


Figure 2.2 – AutoDock Lamarckian genetic algorithms procedure (extracted from the original paper by Morris et al. [107]). f(x) in the figure is the fitness function of the algorithm.

The genetic algorithm search starts with the definition of the genotype and the phenotype. Given the fact that each docking pose is an individual, the arrangement of a ligand can be defined by a set of values describing the translation, orientation and conformation of the ligand with respect to the protein: these are the ligand state variables. Each state variable corresponds to a gene, and the ensemble of genes corresponds to the genotype. The atomic coordinates of the ligand, instead, correspond to the phenotype. The total interaction energy of the ligand with the protein, evaluated using the scoring function, corresponds to the fitness of each individual.

In genetic algorithms, random pairs of individuals are mated using a process of crossover, in which new individuals inherit genes from either parent. In addition, some offspring undergo random mutation, in which one gene changes by a random amount. Selection of the offspring of the current generation occurs based on the fitness of each individual: thus, solutions better suited to their environment reproduce, whereas poorer suited ones die.

In AutoDock, the chromosome (or genotype) is composed by a string of real-valued genes: three Cartesian coordinates for the ligand translation; four variables defining a quaternion specifying the ligand orientation; and one real-value for each ligand torsion, in that order. In total, there is a one-to-one mapping from the state variables of the ligand to the genes of the individual’s chromosome.

At each generation, a user-defined fraction of the population undergoes systematic local search. This is performed both on the translation and rotation of the ligand and on the torsions of its bonds by a modified version of the Solis and Wets method [108], that allows the different types of genes to change with different step sizes.

In the Lamarckian genetic algorithm, genotypic mutation plays a different role than it does in traditional genetic algorithms. Traditionally, mutation generates small changes in the coordinates of the ligand's atoms. This allows a local search in the conformational space. In Lamarckian genetic algorithms, instead, the local search plays this role, while the mutation is needed only for its role in replacing alleles that might have disappeared through selection.

The genetic algorithm iterates over generations until one of the termination criteria is met. At the end of each docking, AutoDock reports the fitness, the state variables, and the coordinates of the docked conformation. Moreover, it carries out conformational cluster analysis on the docked conformations to determine which are similar, reporting the clusters ranked by increasing scoring function value.

2.3.2 AutoDock scoring stage

AutoDock is a grid-based docking method. It precomputes an energy grid approximating the protein. Since the interaction energy can then be approximated by calculating the energy between atoms of the ligand and the appropriate grid points, the docking can be accomplished much faster.

The scoring function adopted by AutoDock uses a semiempirical force field. This estimates the energetics of the process of binding of the molecules in a water environment using pairwise terms to evaluate the interaction between the two molecules and an empirical method to estimate the contribution of the surrounding water.

The free energy of binding is considered equal to the difference between the free energy of the complex and the sum of the free energy of ligand and protein in the unbound state. Therefore the force field evaluates the binding in two steps. The first step evaluates the intramolecular energetics of the transition from the unbound states to the conformation that the ligand or protein will adopt in the bound complex. The second step evaluates the intermolecular energetics of combining the ligand and protein in their bound conformations.

The force field includes six pair-wise evaluations (V) and an estimate of the conformational entropy lost upon binding (ΔSconf):

$$\Delta G = \left(V_{bound}^{L-L} - V_{unbound}^{L-L}\right) + \left(V_{bound}^{P-P} - V_{unbound}^{P-P}\right) + \left(V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}\right)$$

In this equation, L refers to the ligand and P refers to the protein. The first two terms are intramolecular energies for the bound and unbound states of the ligand, and the following two terms are intramolecular energies for the bound and unbound states of the protein. The change in intermolecular energy between the bound and unbound states is in the third parentheses. It is assumed that the two molecules are sufficiently distant from one another in the unbound state that $V^{P\text{-}L}_{unbound}$ is zero.

The pairwise atomic terms include evaluations for dispersion/repulsion, hydrogen bonding, electrostatics, and desolvation:

$$V = W_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} \left(S_i V_j + S_j V_i \right) e^{\left(\frac{-r_{ij}^2}{2\sigma^2} \right)}$$

The weighting constants W were optimized to calibrate the empirical free energy based on a set of experimentally characterized complexes. The first term is a typical 6-12 potential for dispersion/repulsion interactions. Parameters A and B were taken from the Amber force field [109]. The second term is a directional H-bond term based on a 10-12 potential [110]. The parameters C and D are assigned to give a maximal well depth of 5 kcal/mol at 1.9 Å for OH and NH, and a depth of 1 kcal/mol at 2.5 Å for SH. Directionality of the hydrogen bond interaction E(t) is dependent on the angle t away from ideal bonding geometry. Electrostatic interactions are evaluated with a Coulomb potential where q are the charges and ε is the dielectric constant. The final term is a desolvation potential based on the volume (V) of the atoms surrounding a given atom, weighted by a solvation parameter (S) and an exponential term based on the distance. The distance weighting factor σ is set to 3.5 Å.

The force field is calibrated for a united atom model, which explicitly includes heavy atoms and polar hydrogen atoms. Intramolecular energies are calculated for all pairs of atoms within the ligand (or protein, if it has free torsional degrees of freedom), excluding 1-2, 1-3, and 1-4 interactions.

The term for the loss of torsional entropy upon binding (ΔS_{conf}) is directly proportional to the number of rotatable bonds in the molecule (N_{tors}):

$$\Delta S_{conf} = W_{conf} N_{tors}$$

The number of rotatable bonds include all torsional degrees of freedom.

2.3.3 Set up of docking simulations

The docking simulations were prepared with the following steps:

1. preparation of the structure of the ligand and selection of the torsions;
2. preparation of the structure of the receptor;
3. definition of the energy grid;
4. set up of the sampling and cluster analysis parameters.

Preparation of the structure of the ligand

All the ligands used in the ligand-docking calculations for this thesis were taken from the CCDC/Astex Test Set [111], thus they were already minimized using the force field implemented in Sybyl – MAXIMIN module [112]. Gasteiger charges [113] were added to the ligands and the rotatable bonds were identified.

Preparation of the structure of the receptor

Polar hydrogens were added to the pdb structures of each receptor using AutoDockTools. Gasteiger charges were added and the histidines were set as neutral, since the docking process was simulated at pH=7.

Definition of the energy grid

For each protein, a grid was defined to allow the evaluation of the receptor energetic contribution. The grid defines also the part of the protein that is accessible for the conformational search of the ligand.

The following parameters were set up to define each grid:

- distance between grid point;
- center coordinates;
- number of points for each dimension.

The default value of distance between grid points was chosen (0.375 Å). The coordinates of the center and the size of the grid were separately chosen on each single protein structure, in order to contain the whole binding site.

Set up of the sampling and cluster analysis parameters

The AutoDock parameters used during the sampling are reported in Table 2.2.

The values of two parameters are in particular crucial for a thorough sampling of the conformational space: the number of energy evaluations (ga_num_evals) and the number of generations (ga_num_generations), because they determine when the Lamarckian genetic algorithm generation step should stop. The values selected for these two indices were far above the default parameters, in order to ensure a complete search.

Once the docking poses are generated and the scores are calculated, AutoDock, performs a cluster analysis on the basis of their distance in RMSD. The maximum RMSD value that two poses should have to be in the same cluster was set to 2 Å.

Tab 2.2 – List of the AutoDock parameters and values used in this thesis.

Parameter name	Meaning of the parameter	Value
ga_run	No. LGA runs	100
ga_pop_size	No. of individuals for each LGA generation	150
ga_num_evals	Maximum no. of energy evaluations for each LGA run	25,000,000
ga_num_generations	Maximum no. of generations for each LGA run	27,000
ga_elitism	No. of best individuals retained for the next generation	1
ga_mutation_rate	Fraction of genetic mutation	0.02
ga_crossover_rate	Fraction of crossover	0.8
ga_cauchy_alpha	α parameter for Cauchy's distribution	0.0
ga_cauchy_beta	β parameter for Cauchy's distribution	1.0
ls_search_freq	Fraction of individuals undergoing local search for each LGA generation	0.06
tran0	Initial translational coordinates (Å)	random
quat0	Initial rotational coordinates (°)	random
dihe0	Initial torsional coordinates (°)	random
tstep	Ligand translational step (Å)	0.2
qstep	Ligand rotational step (°)	5.0
dstep	Ligand torsional step (°)	5.0

2.3.4 Docking results evaluation

To evaluate the results obtained by ligand docking into both the experimental structure and homology models of the protein, the dRMSD was calculated between the ligand-site distances in the docked complex and the X-ray ligand-site corresponding distances:

$$dRMSD = \sqrt{\frac{\sum_i \sum_j \left(d_{ij}^{x} - d_{ij}^{m}\right)^2}{N}}$$

where x and m are the experimental and docked complex, respectively, d the vectors of the distances between the ligand and the binding site heavy atoms, i and j the indices of the atoms and N is the number of comparisons performed.

Using this index, the distance calculation takes into account only the deviation on the relative position of the ligand to the residues belonging to the binding site and not, as it is for the RMSD calculation, of the deviation on the absolute position of the ligand in the pose from the crystallographic one. Therefore, when the dRMSD is used for assessing the accuracy of the ligand docking into homology models, the structural differences between the model and the experimental structure are excluded from the evaluation of the quality of docking results.

2.4 Protein-protein docking

Two different protein-protein docking methods were used in this thesis: ZDOCK [114, 115] and HADDOCK [116, 117]. The former uses a rigid-body FFT search algorithm, the latter an energy minimization approach.

2.4.1 ZDOCK and ZRANK

ZDOCK is a protein-protein docking algorithm which generates docking poses optimizing shape complementarity, desolvation and electrostatics by using Fast Fourier Transform (FFT).

Representing the proteins in a simplified way and treating the molecules as rigid bodies, it allows the generation of thousands of candidate poses in a relatively short time. This allows an almost complete sampling of the intermolecular conformational space of the complex, performed through a systematic search of the rotational and translational space for the ligand relative to the receptor, which remains fixed at its starting orientation. For the rotational search, evenly distributed Euler angles (usually Δ=6°) are used. For each rotation, the algorithm scans the translational space using FFT based on a three-dimensional grid with step size of 1.2 Å.

In ZDOCK, in fact, each interacting protein is mapped to a 3D grid of NxNxN points. To the cells of the grid are assigned appropriate values representing features of the part of the protein mapped such as desolvation parameters, charges or values representing surface exposure (see for example Figure 2.3). A separate 3D function composed by a real and imaginary part is needed to represent each of these physical parameters, but they can be linearly combined to form an overall score, that is what is optimized during the sampling stage of ZDOCK.

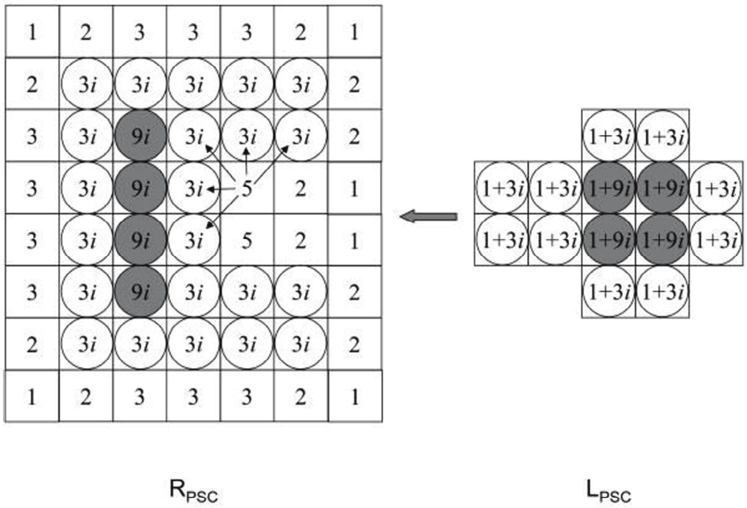


Figure 2.3 – The ZDOCK schema for the association of Pairwise Shape Complementarity function values to the grid cells representing the protein surface (extracted from the original paper by Chen et al. [115]). Protein atoms are indicated using circles, with open circles indicating surface atoms and shaded circles indicating core atoms. For clarity, here the grid spacing equals atom diameter. Grid points whose values are 0 have been omitted from the figure. The value assigned to each grid point is indicated, with i = $\sqrt{-1}$. For each grid point in the open space of RSC (not corresponding to any atom), the number of atoms within a distance cutoff is recorded. Small arrows point out the atoms that are within the distance cutoff of a grid and thus contribute to its score.

The 3D functions used for the scoring are:

- Pairwise shape complementarity (PSC) [115]:
  for each grid cell, it records the position with respect to the surface of the protein, thus it discriminates between the core residues, the surface residues and the solvent (distinguishing between space far off or near the protein surface).
- Desolvation (IFACE) [118]:
  this derives from statistical potentials of desolvation. It takes into account the atom type of the atom comprised in the grid cell, considering 18 atom types.
- Electrostatics (ELEC) [119]:
  the electrostatic contribution to the docking score is calculated based on the Coulombic formula and is expressed as the correlation betwen the partial charges of the ligand atoms with the electrostatic potential due to the receptor atoms.

In this thesis, the poses generated by ZDOCK where scored by using ZRANK [120], a scoring function that includes terms of van der Waals, electrostatics and desolvation in a weighted linear sum:

$$Score = w_{vdW} E_{vdW} + w_{elec\_lr} E_{elec\_lr} + w_{elec\_sr} E_{elec\_sr} + w_{ds} E_{ds}$$

where elec_sr is for ‘short-range electrostatics’ and elec_lr for ‘long-range electrostatics’. The van der Waals contribute is estimated using the Lennard-Jones 6-12 potential from the CHARMM19 polar hydrogen potential [121]. This is calculated for all atoms at an interatomic distance < 8.0 Å. For electrostatics interactions, the Coulomb equation is used, with a 1/r distance-dependent dielectric. For short-range electrostatics (distance < 5.0 Å), partial charges from the CHARMM 19 polar hydrogen potential are used too. For electrostatics interactions at distances greater than 5.0 Å, only fully charged side chain atoms are used, with charges assigned as in Gray et al. [122]. The fact that the short-range electrostatics is determined using polar hydrogen partial charges allows for the hydrogen bonding and polar forces to be calculated within the electrostatics. Pairwise Atomic Contact Energy (ACE) [123] is used to calculate the desolvation energy.

2.4.2 HADDOCK

HADDOCK [116, 117] is a docking method which makes use of hints about protein interfaces to drive its sampling stage. This information can come from several different sources, both experimental and theoretical [42, 43].

The residues that are believed to be involved in the interaction are defined, within the program, as ‘active residues’; ‘passive residues’ are their solvent-accessible neighbours. Active and passive residues are used to define a network of ambiguous interaction restraints (AIRs) between the molecules to be docked. An AIR defines that a residue on the surface of a biomolecule should be in close vicinity to another residue (or group of residues) on the partner biomolecule when they form the complex. This is expressed as an ambiguous distance restraint between all atoms of the source residue to all atoms of all target residue(s) that are assumed to be in the interface of the complex (Figure 2.4). The effective distance between all those atoms, d_{iAB}^{eff}, is calculated as follows:

$$d_{iAB}^{eff} = \left(\sum_{m_{iA}=1}^{N_{Aatom}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{Batom}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-1/6}$$

where N_{Aatom} indicates all atoms of the source residue on molecule A, N_{resB} the residues defined as interface of the target molecule B, and N_{Batom} indicates all atoms of a residue on molecule B. The $1/r^6$ sum averaging mimics the attractive part of a Lennard-Jones potential and ensures that the AIRs are satisfied as soon as any two atoms of the biomolecules are in contact. The AIRs are incorporated as an additional energy term to the energy function that is minimized during the docking.

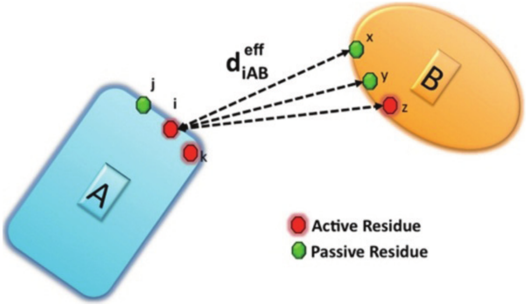


Figure 2.4 – Illustration of AIRs used in HADDOCK to drive docking (extracted from the paper by Karaca et al. [124]). Active residues correspond to residues experimentally identified or predicted to be at the interface. Passive residues are surface neighbors of active residues. AIRs are defined for each active residue with the effective distance being calculated from the sum of all individual distances between any atom of an active residue and any atom of all active and passive residues on the partner molecule.

The ambiguous nature of these restraints allows the use of experimental data that often provide evidence for a residue being at the interface as a driving force for the docking. In fact, the AIRs define a network of restraints without specifying the relative orientation of the molecules, minimizing the necessary search through conformational space needed to assemble the interfaces and allowing a thorough sampling of the limited conformational space defined by the constraints.

The workflow of HADDOCK consists in three steps:

- randomization of orientations and rigid-body energy minimization (EM) driven by interaction restraints (it0);
- semi-flexible simulated annealing in torsion angle space in which side-chains and backbone atoms at the interface are allowed to move (it1);
- final refinement in Cartesian space with explicit solvent (typically water).

In the last two stages, flexible segments are typically defined automatically based on the identified intermolecular contacts.

Randomization of orientations and rigid-body energy minimization

In the randomization stage, the two molecules are positioned at 150 Å from each other and each protein is randomly rotated around its centre-of-mass.

After this, rigid body energy minimization is performed: first, four cycles of orientational optimization are performed in which each protein in turn is allowed to rotate to minimize the intermolecular energy function. Then both translations and rotations are allowed, and the two proteins are docked by rigid body EM. The number of docking poses to be generated at this step is user-defined. Typically, 10000 or 5000 solutions are produced and the best 400 or 200 solutions in terms of intermolecular energies are then refined.

Semi-flexible simulated annealing in torsion angle space

The second stage consists of three simulated annealing refinements. In the first one (1000 steps from 2000 to 50 K, 8 fs time steps), the two proteins are considered as rigid bodies and their respective orientation is optimized. In the second simulated annealing (4000 steps from 2000 to 50 K, 4 fs time steps), the side-chains at the interface are allowed to move. In the third simulated annealing (1000 steps from 500 to 50 K, 2 fs time steps), both the side-chains and backbone at the interface are allowed to move. This permits some conformational rearrangements at the interface. The resulting poses are then subjected to 200 steps of steepest descent EM.

Final refinement with explicit solvent

The final stage consists of a mild refinement in an 8 Å shell of TIP3P water molecules [125]. A 2 fs time step is used for the integration of the equation of motions. The system is first heated to 300 K (500 steps at 100, 200, and 300 K) with position restraints (kpos = 5 kcal mol^{-1} A^{-2}) on all atoms except for the flexible side chains at the interface. MD steps (5000) are then performed at 300 K with position restraints only on non-interface heavy atoms (kpos = 1 kcal mol^{-1} A^{-2}). During the final cooling stage (1000 MD steps at 300, 200, and 100 K), the position restraints are limited to backbone atoms outside the interface.

The final structures are usually clustered by HADDOCK with the algorithm described in Daura et al. [126], using a cut-off of 7.5 Å RMSD, only counting the interface backbone atoms of the ligand. The resulting clusters are ranked according to their average interaction energies.

During the different stages of the docking protocols, solutions are scored as follows:

$$it0: \quad score = 0.01 * E_{vdW} + 1.0 * E_{elec} + 0.01 * E_{AIR} - 0.01 * BSA + 1.0 * E_{desolv}$$
$$it1: \quad score = 1.0 * E_{vdW} + 1.0 * E_{elec} + 0.1 * E_{AIR} - 0.01 * BSA + 1.0 * E_{desolv}$$
$$water: \quad score = 1.0 * E_{vdW} + 0.2 * E_{elec} + 0.1 * E_{AIR} + 1.0 * E_{desolv}$$

E_{AIR} represents the (ambiguous interaction) distance restraint energy; BSA is the buried surface area; E_{vdW} and E_{elec} are the van der Walls and electrostatics energies calculated with an 8.5 Å distance cut-off using the OPLS united atom force field [127] and E_{desolv} the desolvation energy calculated using the parameters of Fernandez-Recio et al. [128].

2.4.3 Indices to evaluate protein-protein docking results

In this thesis, the protein-protein docking results presented in Chapter 3 were evaluated by comparison with the experimental structure of the complex, according to the CAPRI criteria [129]. These take into account three parameters:

- l-RMSD: RMSD on the ligand backbone, after superimposition of the receptors of the docked and the native complex;
- i-RMSD: RMSD on the interface backbone, after superimposition of the interface of the docked and the native complex;
- fnat: fraction of the native contacts found in the docking solution.

The interface backbone was defined as the backbone atoms of the residues within 10 Å of any heavy atom of the interacting partner; fnat was calculated at a distance of 5 Å.

Thus, both the reproduction of the overall geometry and interactions of the complex are used in evaluating docking results.

According to the values adopted by these parameters, the docking solutions are classified in three categories, as reported in Table 2.3. The three-star solutions are those having the highest accuracy, then come the two stars (medium accuracy) and the one star (acceptable accuracy). The solutions that do not fall in these three categories are classified as inacceptable.

The protein-protein docking results presented in Chapter 5 and obtained by docking one protein into both the experimental structure and the homology models of the other protein, instead, were evaluated only on the basis of their i-RMSD with respect to the native complex.

Tab 2.3 – CAPRI rules for the assessment of protein-protein docking results.

Prediction category	l-RMSD (Å)	i-RMSD (Å)[a]	fnat[b]
High accuracy (★★★)	≤ 1.0	≤ 1.0	≥ 0.5
Medium accuracy (★★)	≤ 5.0	≤ 2.0	≥ 0.3
Acceptable accuracy (★)	≤ 10.0	≤ 4.0	≥ 0.1
Incorrect	> 10.0	> 4.0	< 0.1

a) Interface residues are those within 10 Å of the interacting partner.

b) Calculated on residues within 5 Å of the interacting partner.

Chapter 3

Combining different sampling strategies for protein-protein docking

This chapter illustrates the development and testing of ZADDOCK, a new protein-protein docking approach which is a combination of two of the best performing protein docking methods that make use of different sampling strategies: rigid-body *ab initio* search (ZDOCK) and EM/MD sampling (HADDOCK).

Through an analysis of the performance of this new method on a wide and representative set of protein-protein complexes corresponding to different types of interaction, we demonstrated that ZADDOCK can be a reliable and useful tool, not only for generating possible binding modes, but also for providing accurate information on the interactions that take place in protein complexes and confirmed the potentiality of combining different docking methods to overcome their limitations in sampling and sum their strengths.

3.1 Introducing ZADDOCK, a new protein-protein docking method that uses FFT search for interface prediction

As discussed in the introduction chapter, the search stage in protein-protein docking is a crucial step, but it is constrained by the current computational limits. Nowadays, in fact, it is not possible to perform a complete search in the conformational space of the interacting partners, including a full flexible treatment of the partner molecules. Consequently, every docking method is affected by approximations introduced in order to make the sampling stage more efficient. Over the years, various strategies have been developed to solve the sampling problem. Among them, the most popular approaches are: fast Fourier transform (FFT), Monte Carlo (MC) search and energy minimization/molecular dynamics (EM/MD). FFT-based methods [130] generate thousands of conformations of the complex in a relatively short time, thus allowing an almost complete search of the relative orientations in the conformational space. Since they are very fast, they can be used even in the total absence of information regarding the interaction patches. These methods have, however, the disadvantage of treating the proteins as rigid bodies.

MC and EM/MD-based methods, on the other hand, can take into account the flexibility of the individual partners. The limit of these approaches is that the sampling requires a great amount of computational time. For this reason, the docking methods that implement this kind of sampling usually need restraints to limit both the search of the relative position of the partners and the flexible search. Restraints can be defined based on several sources, among which biophysical, biochemical and bioinformatics data [42, 43].

HADDOCK [116, 117] (for a detailed description of the method see Paragraph 2.4.2) belongs to the second class of docking methods: it is based on a combination of rigid-body EM-driven docking and flexible molecular dynamics refinement stages and relies strongly on external information. The latter is both its main strength and weakness: it can achieve high-accuracy results provided that sufficient data are available, but can also easily be misled by wrong information [131]. In the absence of experimental information, bioinformatics methods can be used to predict putative interface patches. This can be done either using traditional interface prediction tools [132] that consider residues conservation, interface propensities and physico-chemical characteristics of the amino-acids exposed to the solvent, or by exploiting docking methods themselves [128, 133]. In particular, this second strategy seems very promising, since it considers both proteins and thus can be guided both by the physics of the interaction and by the complementarity in the physico-chemical features of their surface and shape.

To explore the potentiality of this latest approach and allow the use of HADDOCK in absence of any experimental information, we combined HADDOCK with ZDOCK [114, 115], an FFT docking software that does not require any input data, except the structures of the interacting partners (for a detailed description of the method see Paragraph 2.4.1). This integrated approach, ZADDOCK, allows to overcome the weaknesses of the isolated approaches, combining in a synergistic manner their strengths to achieve better results. In particular, the fast and almost complete rigid-body search of ZDOCK is used to sample the relative orientations of the two interacting partners and to predict the interface patches. The resulting initial binding modes are subsequently subjected to the flexible refinement of HADDOCK to obtain high quality predictions. As a results, the main goals achieved by ZADDOCK are:

- the possibility of using the flexible refinement stage of HADDOCK without the need of any experimental data to guide the sampling step. This feature makes ZADDOCK a powerful tool to study complexes for which no experimental information is available and bioinformatics interface prediction fails;
- the improvement in the description of the interactions occurring inside protein complexes, a key information to drive subsequent experimental work.

The strategy of combining different docking methods with the aim of improving the final results has been proposed also in some recent papers [44, 62-65] indicating that this is indeed a very promising avenue to improve docking results. Pierce and Weng [64], in particular, proposed a strategy for combining the initial search stage of ZDOCK with the structural refinement of another docking methods, RosettaDock [122], which implements side-chain repacking and a Monte Carlo search at the interface of the ligand. The combination of those two methods, together with the development of a scoring function *ad hoc*, did lead to very good docking results, improving the performance of ZDOCK in terms of geometrical interface adherence of the docking poses to the native structures of the complexes.

3.2 Computational approach: integration of ZDOCK and HADDOCK into ZADDOCK

Two existing docking programs, ZDOCK 3.0 [118] and HADDOCK 2.1 [134], were combined in a streamlined and automated procedure into ZADDOCK. The ZADDOCK workflow is illustrated in Figure 3.1: the first two steps of ZADDOCK consist in a standard ZDOCK+ZRANK procedure, while the last two steps coincide with the semi-flexible simulated annealing (*it1*) and the final refinement stage (*water*) of HADDOCK (see Paragraph 2.4 for a thorough explanation of the two docking approaches).

After having built topologies and missing atoms for the starting structures, the initial rigid body docking stage is performed by FFT search using ZDOCK with a rotational sampling of 6° and a grid resolution of 1.2 Å (corresponding to the fine sampling settings in ZDOCK). The top 10000 docking solutions are written to disk. This step is followed by a re-ranking stage, during which the top 10000 solutions generated by ZDOCK are ranked using ZRANK [120]. The 400 best-scored solutions are then analyzed: for each of them, the contacts between the Cα atoms of the two proteins that are within a 10 Å cut-off ('interface Cα') are extracted and a list of HADDOCK restraints is created. The distance restraints are set to the measured Cα-Cα distance ± 1.0 Å error bound. These restraints are then used in the two subsequent semi-flexible refinement steps. During this refinement, 50% of the restraints are randomly discarded for each docking model, to allow some sampling around the starting orientations. In order to deal with the severe clashes frequently found in FFT-based docking models, the original HADDOCK refinement protocol was modified: the Lennard-Jones potential was truncated at 1.5 Å interatomic distance and then progressively increased to 0.5 Å (the default value in HADDOCK) during the simulated annealing stage in it1. Moreover, surface and centre-of-mass restraints were tightened, to avoid a separation of the interacting molecules due to the high energy caused by the clashes.

The final solutions after refinement in explicit solvent (water stage) are then ranked based on the HADDOCK score defined as:

$$HADDOCKscore = 1.0 * E_{vdW} + 0.2 * E_{elec} + 0.1 * E_{AIR} + 1.0 * E_{desolv}$$

(see Paragraph 2.4.2 for a detailed description of the score).

The final predictions are also subjected to cluster analysis, as implemented in the standard HADDOCK protocol (see Paragraph 2.4.2), using a cut-off of 7.5 Å RMSD, only counting the interface backbone atoms of the ligand. For the analysis reported in this work, the resulting clusters were ranked according to the average interaction energies of their four best-scored predictions.

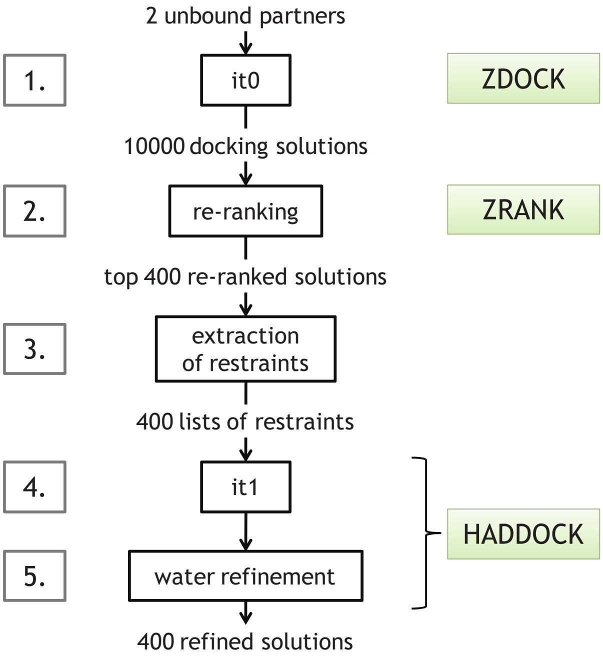


Figure 3.1 – ZADDOCK workflow

3.3 ZADDOCK performance

3.3.1 Test set

The performance of ZADDOCK was tested on the protein-protein docking benchmark 3.0 [135]. This benchmark set is composed of 124 protein-protein complexes of various types divided into four groups:

- 13 antibody/antigen unbound (abbreviation: A);
- 12 antibody/antigen bound (AB);
- 35 enzyme/inhibitor (E);
- 64 other (O).

They are also classified by their level of difficulty [136]:

- 88 rigid-body (r);
- 19 medium difficulty (m);
- 17 difficult (d).

The simplest cases are targets for which none of the components undergoes significant conformational changes upon binding, whereas the most challenging are the ones for which the binding causes large conformational changes. The percentage of rigid-body, medium difficulty and difficult cases for each class of complexes is reported in Table 3.1. For all cases, the docking was performed starting from the unbound structures when available, and the protonation state of histidines was defined using the WHATIF webserver [137].

The docking solutions were evaluated according to the CAPRI criteria [129], as described in Paragraph 2.4.3. Superimposition of the structures and RMSD calculations were performed using the McLachlan algorithm [86] as implemented in the program ProFit.

Table 3.1 – Percentage of rigid-body, medium difficulty and difficult cases present in each class of complexes of the test set.

Class	% of rigid-body cases	% of medium difficulty cases	% of difficult cases
E	80%	14%	6%
O	59%	20%	20%
A	85%	8%	8%
AB	92%	0%	8%

3.3.2 ZADDOCK compared to HADDOCK *ab initio* and CPORT+HADDOCK

As one of the goals of developing ZADDOCK was to allow the use of HADDOCK in the absence of any experimental data or bioinformatics interface predictions, a first comparison was made between the performance of ZADDOCK and HADDOCK: 1) run without experimental restraints ('*ab initio*') and 2) guided by interface prediction data obtained by CPORT [138].

The HADDOCK *ab initio* mode consists in performing the docking calculations with centre-of-mass restraints corresponding to distance restraints defined between the geometric centres of the Cα atoms of the various partners. The distance is automatically defined based on the size of each molecule. This procedure allows to force contacts between the two molecules, without defining specific restraints. CPORT [138] is a meta predictor for protein-protein interfaces that combines predictions from up to six different servers. The resulting predictions, tuned for high sensitivity (coverage of the true interface), can be used to define active and passive residues in HADDOCK.

ZADDOCK results for 'enzyme/inhibitor' and 'others' docking cases of the benchmark 2.0 [139] were compared to the corresponding cases obtained with either *ab initio* or CPORT-driven docking with HADDOCK [138]. The results are summarized in Figure 3.2, where the histograms report the percentage of cases of the benchmark showing at least an acceptable solution (one star or better, Fig. 3.2a) or a medium accuracy solution (two stars or better, Fig. 3.2b) among the predictions. From the histograms, it is clear that ZADDOCK outperforms by far HADDOCK *ab initio*, with the highest difference in the percentage of acceptable solutions in the top100 (33%) (Fig. 3.2a). This is an expected result since centre-of-mass restraints do not contain much information to guide the docking and a more thorough sampling of the interaction space would be required: HADDOCK was clearly not developed as an *ab initio* docking program.

The differences between ZADDOCK and CPORT+HADDOCK are less extreme, with ZADDOCK outperforming CPORT+HADDOCK in all cases except for the number medium and high accuracy solutions at the top (Fig. 3.2b): CPORT+HADDOCK ranks a two- or three-star solution at the top in 8% of the cases, while this percentage for ZADDOCK is 5%. These results are rather unexpected, since they mean that the combination of a mainly geometric docking method with an energy-based flexible refinement achieves better results than when using interface predictions.

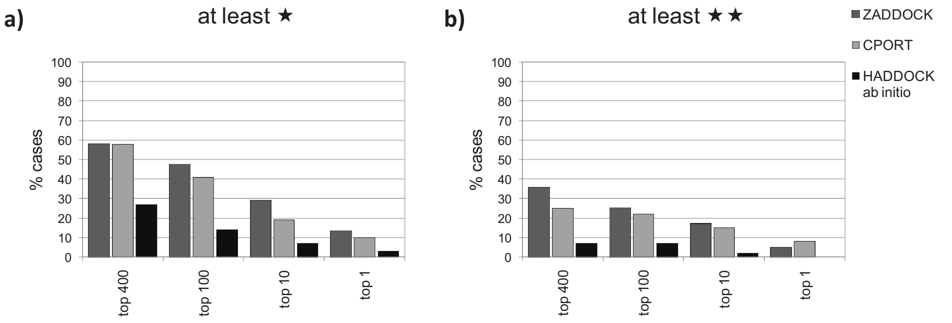


Figure 3.2 – Comparison of the performance of ZADDOCK, HADDOCK *ab initio* and CPORT+HADDOCK. Percentage of cases of the benchmark for which at least one structure of: a) one star (or better) quality or b) two star or better quality was obtained among the whole set of docking solutions (top 400) or the best 100, 10 or 1.

3.3.3 ZADDOCK results on different classes of complexes

To determine the generality and versatility of ZADDOCK, its performance was evaluated separately for the different classes and difficulties of the complexes in the benchmark set. The histograms in Figure 3.3 report the percentage of cases of the various complex classes for which at least a one-star (or better) solution (Fig. 3.3a) or a two-star (or better) solution (Fig. 3.3b) was found among the final set of refined predictions.

For the enzyme/inhibitor (E), antibody/antigen unbound (A) and antibody/antigen bound (AB) classes, the percentage of one-star solutions is above 65%. This value decreases for the top 100 and top 10 solutions. One can also observe a clear difference in performance between the various classes of complexes, with the enzyme/inhibitor complexes performing best. The other two categories have values above 30% both for the top 100 and the top 10 predictions. When considering the top ranked solutions, the three classes show again about the same performance. Interestingly, for the one-star (or better) predictions ZADDOCK showed nearly the same overall performance when evaluated over the AB and A classes. This indicates that the method is not very sensitive to the conformation of the starting structure (bound or unbound). Good results were achieved also for the most difficult category, ‘others’, that is characterized by a very high degree of heterogeneity. For the two-star (or better) solutions, the same scenario can be observed: the best performance was achieved for the E class, followed by the AB, A and O categories.

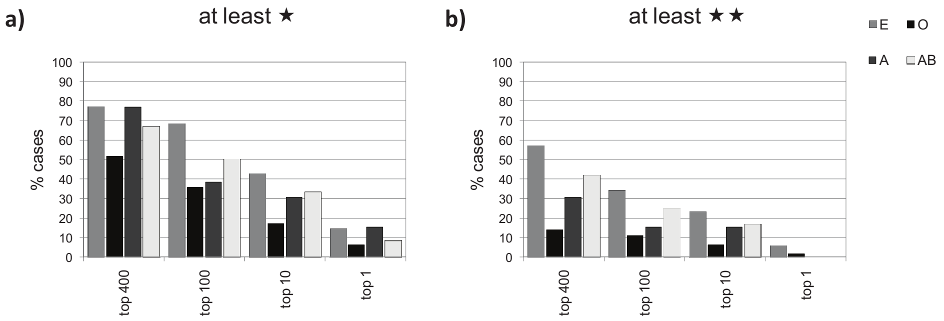


Figure 3.3 – ZADDOCK performance on the classes of complexes of the benchmark (E = enzyme/inhibitor, O = others, A = antibody/antigen unbound, AB = antibody/antigen bound). The histograms report the percentage of cases of the benchmark for which at least one structure of: a) one star (or better) quality or b) two star or better quality was obtained among the whole set of docking solutions (top 400) or the best 100, 10 or 1.

Note that the differences between classes also reflect to some extent their distribution between easy, medium and challenging cases. A similar performance for rigid-body and medium-difficulty cases is found when considering the percentage of one-star (or better) predictions among the top 400 (Fig. 3.4a and 3.4b). When considering only the top 100 or top 10, better results are found for the rigid-body cases. The difference in performance for these two classes becomes negligible when considering only the best-ranked pose. This means that the sampling stage of ZADDOCK is able to achieve similar results for the complexes for which minor or medium changes occur upon binding. Scoring, on the other hand, fails to recognize many near-native structures in the top 100 and 10 for the medium-difficulty cases, penalizing thus the cases for which the starting structures differ more from the bound form. As for the difficult complexes (RMSD > 2.2 Å for the interface Cα atoms between unbound and bound conformation), in only about 10% of the cases is a one-star solution present among the final refined predictions. This is not surprising since rigid-body docking is unable to sample large conformational changes that occur upon binding and the semi-flexible refinement stage in HADDOCK can only lead to rather small rearrangements (typically up to a maximum of 2 Å). Clearly, if the rigid-body initial search does not sample near-native conformations, the refinement stage is unable to generate medium-accuracy predictions.

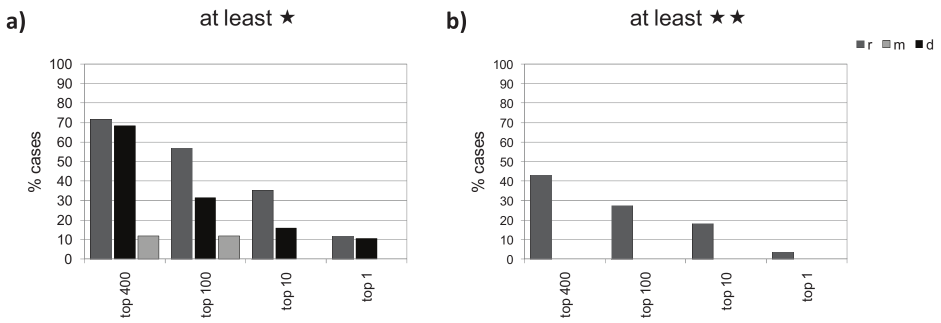


Figure 3.4 – ZADDOCK performance on the classes of difficulty of the benchmark (r = rigid-body, m = medium difficulty, d = difficult). The histograms report the percentage of cases of the benchmark for which at least one structure of: a) one star (or better) quality or b) two star or better quality was obtained among the whole set of docking solutions (top 400) or the best 100, 10 or 1.

In summary, this analysis indicates that ZADDOCK performs homogeneously well for all the categories of complexes in the benchmark set depending on their degree of difficulty (conformational changes).

3.3.4 ZADDOCK compared to ZDOCK+ZRANK

To assess the effect of the flexible refinement on the overall performance of the method, the results of the ZDOCK+ZRANK stage and of the whole ZADDOCK process were compared (Fig. 3.5).

During the refinement, no significant changes occurred in the number of test cases for which a one-star solution is found among the entire set of predictions. Considering the top 100 predictions, the HADDOCK scoring function seems slightly better than ZRANK in detecting the one-star solutions. Considering the two- and three-star solutions, ZDOCK+ZRANK performs a bit better with a 35% success percentage compared to 31% for ZADDOCK. It appears thus that a number of good (two-stars or higher) solutions are lost during the refinement step. This can be attributed to clashes in the initial rigid-body pose that were removed in the flexible refinement stage at the cost of the accuracy of the model (see below).

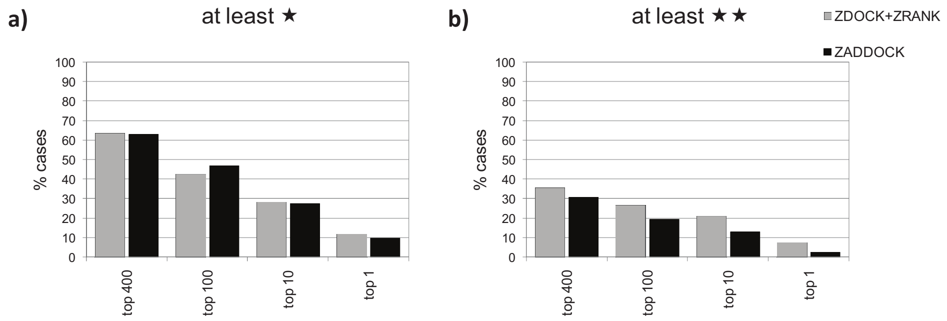


Figure 3.5 – Comparison of the performance of ZADDOCK and ZDOCK+ZRANK. Percentage of cases of the benchmark for which at least one structure of: a) one star (or better) quality or b) two star or better quality was obtained among the whole set of docking solutions (top 400) or the best 100, 10 or 1.

The rank of the first acceptable model (one star) in the top 400 predictions was also analyzed (Table 3.2) to compare the two methods. A better ranking of the best-scored pose implies, in fact, a higher probability of detecting a near-native prediction among the docking poses. For 79 cases of the benchmark at least one acceptable model was present in the top 400 docking poses; for one case (2HMI), the acceptable model was lost after refinement. In 53% of the successful cases the flexible refinement leads to an improvement in ranking, while the initial ZRANK scoring performed better in 37% of the cases (and 10% did not change).

Note however that the classical scoring in HADDOCK is performed on a cluster basis and not individual structures. The standard HADDOCK cluster analysis was performed on the final ZADDOCK solutions, ranking the clusters obtained on the basis of the average score of the four best-scored poses included in each cluster. The results highlighted that in 64% of the cases which presented at least an acceptable solution among the top 400 predictions at least one cluster could be found that contained a star pose among the four best scored ('star-cluster') (some statistics on the best-ranked star-clusters are reported in Table 3.3). Moreover, 38% of those 'star-clusters' were ranked as first.

Table 3.2 – Rank of the first-ranked star solutions for ZDOCK+ZRANK and ZADDOCK for the cases for which acceptable (one star) models were found in the top 400 solutions. Cases for which ZADDOCK improved the rank are indicated in bold.

PDBID	rank ZDOCK+ZRANK	rank ZADDOCK	PDBID	rank ZDOCK+ZRANK	rank ZADDOCK
1A2K	288	**30**	1NCA	20	255
1ACB	29	**9**	1NW9	324	**162**
1AHW	6	**2**	1OPH	237	**44**
1AVX	2	**1**	1PPE	1	1
1AY7	17	**4**	1QFW_HL	1	3
1B6C	1	1	1R0R	25	77
1BJ1	1	2	1RLB	13	26
1BUH	144	**64**	1S1Q	242	302
1BVK	1	87	1TMQ	154	**26**
1BVN	1	2	1UDI	180	323
1CGI	4	**2**	1VFB	117	**111**
1DFJ	1	1	1WEJ	4	**1**
1E6E	2	2	1WQ1	53	**1**
1E6J	28	**8**	1XD3	2	8
1E96	102	107	1XQS	13	**1**
1EAW	2	2	1YVB	1	2
1EER	151	**35**	1Z0K	6	115
1EFN	27	144	1Z5Y	52	**21**
1EWY	8	35	1ZHI	40	**17**
1F34	120	**108**	2B42	7	**4**
1F51	203	**86**	2BTF	16	**4**
1FSK	1	2	2CFH	143	**42**
1GP2	261	**156**	2FD6	39	101
1GPW	3	**2**	2H7V	263	**214**
1GRN	342	**128**	2HLE	1	4
1HE1	50	**22**	2HMI	280	-
1I2M	132	**80**	2HRK	167	310
1I9R	22	54	2I25	206	**175**
1IQD	1	1	2JEL	142	**64**
1J2J	3	8	2MTA	66	**54**
1JPS	13	**1**	2NZ8	277	**149**
1K4C	352	**244**	2O8V	3	8
1K5D	277	**221**	2OT3	76	**66**
1K74	1	3	2PCC	6	13
1KXP	7	**2**	2SIC	1	1
1MAH	2	**1**	2SNI	12	19
1ML0	1	1	2UUY	11	16
1MLC	39	120	2VIS	292	320
1N2C	292	**20**	7CEI	3	4
1N8O	54	**12**			

Table 3.3 – Statistics for the best-ranked ‘star clusters’ for the cases for which clusters containing acceptable (one star) models among the four best-scored elements were found. (*Continues*)

PDBID	cluster rank[a]	HADDOCK score[b]	no. of structures	average iRMSD[b]	average fnat[b]
1ACB	5	125.27	7	4.18 ± 0.70	0.28 ± 0.09
1AHW	12	635.20	15	3.16 ± 0.40	0.45 ± 0.10
1AVX	1	-160.85	58	3.21 ± 0.95	0.58 ± 0.17
1AY7	11	-30.50	6	4.06 ± 0.34	0.49 ± 0.06
1B6C	1	-217.43	79	2.73 ± 0.25	0.77 ± 0.07
1BJ1	1	-86.52	29	2.08 ± 0.15	0.78 ± 0.05
1BVN	1	-174.04	60	1.90 ± 0.23	0.71 ± 0.06
1CGI	1	-163.56	64	3.68 ± 0.41	0.45 ± 0.04
1DFJ	1	-288.81	34	2.80 ± 0.15	0.65 ± 0.06
1E6E	1	-149.80	73	2.06 ± 0.14	0.50 ± 0.07
1E6J	4	27.45	25	2.26 ± 0.14	0.73 ± 0.14
1E96	18	250.35	6	3.58 ± 0.33	0.42 ± 0.09
1EAW	1	-161.68	38	1.98 ± 0.15	0.71 ± 0.06
1EFN	19	166.65	6	4.13 ± 0.28	0.30 ± 0.01
1EWY	7	24.09	17	3.71 ± 0.77	0.36 ± 0.14
1F34	27	825.05	5	3.53 ± 0.81	0.31 ± 0.08
1F51	12	267.14	12	4.51 ± 0.65	0.21 ± 0.04
1FSK	1	-141.60	71	1.94 ± 0.15	0.78 ± 0.03
1GPW	1	-85.22	23	3.27 ± 0.77	0.52 ± 0.17
1HE1	10	43.46	7	3.76 ± 0.71	0.39 ± 0.07
1IQD	1	-108.01	59	3.87 ± 0.85	0.33 ± 0.03
1J2J	2	-101.82	20	3.95 ± 0.21	0.39 ± 0.07
1JPS	5	377.17	7	2.79 ± 0.16	0.62 ± 0.12
1K74	1	-168.25	72	2.32 ± 0.54	0.64 ± 0.10
1KXP	1	-182.12	30	3.15 ± 0.43	0.41 ± 0.06
1MAH	1	-166.54	56	1.67 ± 0.06	0.81 ± 0.10
1ML0	1	-168.12	82	2.25 ± 0.39	0.77 ± 0.02
1MLC	21	428.84	9	2.29 ± 0.19	0.70 ± 0.05
1N8O	7	54.90	16	2.22 ± 0.05	0.80 ± 0.06
1PPE	1	-165.43	222	2.03 ± 0.48	0.74 ± 0.13
1QFW_HL	1	-146.79	24	4.21 ± 0.58	0.59 ± 0.15
1R0R	11	212.65	5	4.48 ± 1.06	0.33 ± 0.29
1WEJ	3	-44.39	30	1.73 ± 0.09	0.74 ± 0.06
1XD3	9	58.98	8	4.24 ± 0.64	0.23 ± 0.07
1XQS	2	-149.82	18	3.18 ± 0.12	0.55 ± 0.03
1YVB	2	-71.62	23	2.56 ± 0.74	0.36 ± 0.12
1Z0K	13	15.09	12	3.35 ± 0.63	0.28 ± 0.11
1Z5Y	12	55.48	15	3.92 ± 0.43	0.30 ± 0.10
1ZHI	13	331.04	13	2.82 ± 0.45	0.64 ± 0.07
2B42	3	408.91	14	3.19 ± 0.27	0.59 ± 0.05

a) Rank on the basis of the average HADDOCK score of the four best-scored elements of the cluster.

b) Statistics on the four best-scored elements belonging to the cluster.

Table 3.3 – (*Continued*)

PDBID	cluster rank[a]	HADDOCK score[b]	no. of structures	average iRMSD[b]	average fnat[b]
2BTF	4	241.01	11	5.31 ± 1.68	0.25 ± 0.11
2FD6	14	449.47	6	3.30 ± 0.58	0.54 ± 0.16
2H7V	25	783.33	5	4.38 ± 0.40	0.54 ± 0.04
2HLE	1	-198.06	47	3.40 ± 0.75	0.57 ± 0.12
2I25	26	619.25	4	3.12 ± 0.46	0.38 ± 0.09
2PCC	2	-49.20	24	4.58 ± 0.85	0.21 ± 0.11
2SIC	1	-115.21	38	1.68 ± 0.25	0.79 ± 0.05
2SNI	4	84.29	8	2.72 ± 0.37	0.49 ± 0.19
2UUY	20	305.78	15	2.66 ± 0.16	0.70 ± 0.04
7CEI	2	-208.77	60	2.42 ± 0.10	0.82 ± 0.03

a) Rank on the basis of the average HADDOCK score of the four best-scored elements of the cluster.
b) Statistics on the four best-scored elements belonging to the cluster.

3.3.5 Comparison of the structural quality of the docking models

To detect whether the flexible refinement step improved the structures generated during the rigid body docking stage, an analysis of the clashes, i-RMSD and fraction of native contacts was performed. A clash was defined as a contact with a distance ≤ 2 Å between heavy atoms belonging to the two interacting proteins. As reported in Table 3.4, almost all docking predictions coming from ZDOCK+ZRANK contain clashes (96%) and all the cases of the benchmark had, among their solutions, at least one structure with clashes. The average number of clashes per 1000 Å^2 of BSA (buried surface area) is 3.18. After refinement, only 0.2% of the docking poses and 5 cases out of 124 (4%) still have clashes (number of clashes per 1000 Å^2 of BSA is 0.00), indicating that ZADDOCK refinement protocol is successful in improving the quality of the interface of the rigid-body docking predictions.

Table 3.4 – Atom clashes analysis to the top 400 models[a].

docking step	% of structures with clashes (out of 49600)	% of cases with clashes (out of 124)	average number of clashes per 1000 Å^2 BSA
ZDOCK+ZRANK	96%	100%	3.18 ± 2.02
ZADDOCK	0.2%	4%	0.00 ± 0.01

a) A clash is defined as an intermolecular heavy atom – heavy atom distance ≤ 2Å.

As for the interface RMSD and the fraction of native contacts, both the entire set of top 400 solutions and only the models having i-RMSD ≤ 4.0 Å (which will be called, from now on, 'near-native' solutions) were analyzed to assess the structural improvement caused by the refinement stage. The histograms in Fig. 3.6 show that, on average, the i-RMDS values of the entire set improved after refinement for a majority of the case (around 74%). Taking into account only the 'near-native' predictions, only about 23% improved, the remaining solution having increased i-RMSD values after refinement (average i-RMSD difference: 0.38 Å). Given the presence of a high number of structures with clashes, the proteins are slightly pulled apart as a result of the refinement in order to eliminate the bad contacts, which causes the i-RMSD values of the 'near-native' solutions to increase in most cases. Note that fixing the backbone and only refining or repacking side-chains does not allow to remove the clashes.

When considering fnat values, 20% of the cases improved (average fnat improvement of 0.006), 71% were unchanged and only 10% got worse. As for the 'near-native' solutions, the majority of them (about 55%) showed a higher fraction of native contacts after refinement, a minority did not change and about 38% got worse.

The accuracy improvement in reproducing the interactions between the two partners at atomic details is a very valuable achievement, since ZADDOCK is meant to be used for cases for which no experimental information is available. For such cases, being able to correctly predict contacts is of interest to guide wet-lab experiments or to gain a better understanding of the functioning of a complex.

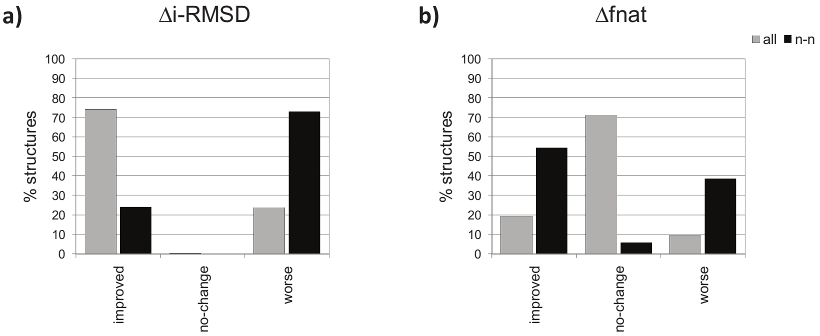


Figure 3.6 – Structural improvement in terms of a) i-RMSD and b) fnat for all the cases of the benchmark, considering the whole set of docking solutions (all) or only the 'near-native' (n-n).

These results are in accordance with what has already been published about the capability of the HADDOCK refinement to change the i-RMSD and fnat values of the rigid-body solutions [117]: while the improvement in i-RMSD values is limited, the fnat of the final predictions is usually very much improved.

3.3.6 Performances of ZADDOCK on an example case

From the above analysis, the typical result of ZADDOCK on a test case, compared to rigid-body docking only, can be generalized as follows:

- better ranking of the best-scored acceptable solution;
- limited (if any) improvement in i-RMSD values;
- improved fraction of native contacts;
- almost no clashes at the interface.

The complex of the Bovine chymotrypsinogen with a Human pancreatic secretory trypsin inhibitor (PDB ID: 1CGI) illustrates nicely these findings. It belongs to the enzyme/inhibitor case, classified as rigid-body in the docking benchmark 3.0. Overall statistics about the number of acceptable solutions, best rank and average number of clashes are reported in Table 3.5 while the structural improvement is given in Table 3.6.

Table 3.5 – Comparison of ZDOCK+ZRANK and ZADDOCK results for the Bovine chymotrypsinogen – Human pancreatic trypsin inhibitor complex (PDBID: 1CGI).

docking step	number of acceptable[a] solutions (top 400)	rank of the first acceptable solution	average number of clashes[b] per 1000 Å^2 BSA
ZDOCK+ZRANK	65	4	3.47 ± 1.97
ZADDOCK	71	2	0.00 ± 0.00

a) (i-RMSD ≤ 4Å or l-RMSD ≤ 10Å) and fnat ≥ 0.1.
b) A clash is defined as an intermolecular heavy atom – heavy atom distance ≤ 2Å.

Table 3.6 – Structure improvement after flexible refinement of the ZDOCK+ZRANK solutions for the 1CGI case.

accuracy parameter	% improved n-n[a] solutions	% no-change n-n[a] solutions	% worse n-n[a] solutions
i-RMSD	18.6	0.0	81.4
fnat	67.4	9.3	23.3

a) n-n = ‘near-native solutions, defined as docking poses with i-RMSD ≤ 4Å.

In Figure 3.7 is reported, as an example, one of the acceptable solutions found for this complex. While no clashes are present in this model, 21 clashes are found in the initial rigid-body solution, which are all relieved after refinement at the cost of the i-RMSD which increases from 2.70 Å to 2.94 Å. At the same time, the fraction of native contacts improves significantly from 0.41 to 0.65. A decrease of the accuracy of the backbone positioning is thus compensated by the plasticity of the side-chains. A detailed view of a few key interacting residues is shown in Figure 3.7, illustrating the improvement after ZADDOCK refinement. In fact, half of the hydrogen bonds of the native complex are found as contacts in the considered ZADDOCK solution and 65% of the non-bonded contacts, even if they don't present the same exact side-chain geometry.

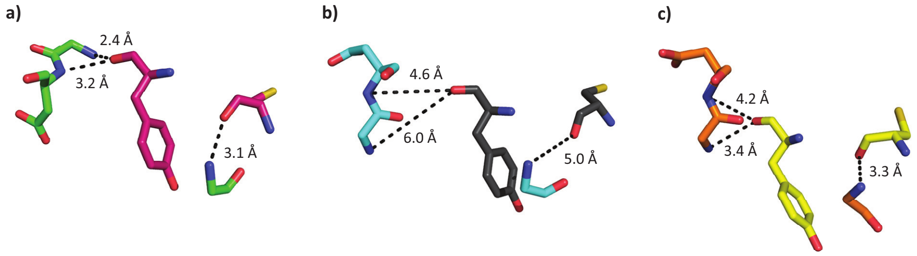


Figure 3.7 – Interatomic distances at the interface of the complex (PDBID: 1CGI). a) native interactions between receptor (green) and ligand (light purple); b) example of one ZDOCK+ZRANK solution (receptor: cyan, ligand: dark grey); c) the corresponding final ZADDOCK model (receptor: orange; ligand: yellow).

3.4 Conclusions

HADDOCK has shown consistent strong performance in CAPRI, belonging to the best performing protein-protein docking methods. As recently discussed, the use of information to guide the docking process is both its strength and weakness [131]. In the absence of any information, HADDOCK has difficulty in selecting near-native solutions for the refinement stage, as demonstrated in a recent paper by de Vries [138]. In order to allow its use even in the absence of any experimental or predicted data, we have developed ZADDOCK, that combines the flexible refinement of HADDOCK with ZDOCK, an FFT docking method whose sampling of the intermolecular conformational space is fast and complete.

Benchmarking ZADDOCK on a representative set of various classes of complexes has indicated an overall good performance provided no major conformational changes are taking place upon binding, a limitation of any rigid-body FFT-based docking approach. The resulting refined models show improved fractions of native contacts with pretty much clash-free interfaces, a significant improvement upon rigid-body only docking.

ZADDOCK will be made available in a future releases of HADDOCK and will also be included into the HADDOCK web server (http://haddock.chem.uu.nl/services/HADDOCK), in order to facilitate its use by a broad structural biology community.

Chapter 4

Low resolution ligand-protein docking: the use of homology models in docking experiments

In this chapter, a large-scale experiment aimed at defining to which extent homology models can be used in ligand-protein docking is described. This was performed on a diverse set including experimental structures and homology models for a group of representative ligand-protein complexes. A wide spectrum of model quality was sampled using templates at different evolutionary distances and several strategies for target-template alignment and modelling. The models obtained were evaluated with a selection of the most used and well-performing model quality indices. The binding geometries were generated using AutoDock, one of the most cited docking programs. In Section 4.1, the background ideas, the proposed computational approach and the results of this study are presented and discussed. Moreover, in Section 4.2 a novel strategy to predict the accuracy of docking results based on indices of model quality is proposed and its reliability is demonstrated for a study case.

4.1 Analysis of the relationships between the accuracy of docking results and the quality of protein models

As discussed in the introduction, homology modelling techniques have made significant contributions at different stages of the drug discovery process, including ligand docking [51, 56, 59], but, given the strong dependence of docking results on the accuracy of the protein structure, the use of good quality models is crucial for such studies.

The problem of identifying the relationship between model quality and docking results accuracy was investigated in some specific cases of high throughput screening [52-54, 56, 140]. However, no clear trend resulted from such studies and, at the present time, there are no general rules for predicting the accuracy of docking results on homology models, even though this would be of great interest both for the docking and the modelling community.

In particular, standard indices to measure the quality of models with reference to the experimental structures [87] could be used for identifying an existing correlation between model quality and docking accuracy, whereas indices of quality assessment, that can be derived without knowledge of the native structure, could provide knowledge-based rules for directly predicting the quality of docking results. Due to the importance and potential of such indices for structure prediction, new methods have been recently developed and a new prediction category, 'model quality assessment', was introduced since the 7th edition of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment [141], but their potential in ligand-protein docking has still to be tested and exploited.

On the basis of the above considerations, the aim of this experiment was to investigate the relationships between the accuracy of ligand docking into homology models and the quality of the models, and to verify if state-of-the-art indices for model quality assessment can be regarded as reliable tools for direct and *a priori* prediction of the accuracy of docking simulations [66].

4.1.1 Computational approach

Reference set of protein-ligand complexes

The test set selected for this work is composed by a reference group of X-ray protein structures and by a corresponding dataset of theoretical models. The reference group was chosen from structures included in the CCDC/Astex Test Set [111], a large and diverse set of known protein-ligand complexes developed for validating ligand-protein docking methods. A subset of cases with resolution of 2.0 Å or better was extracted and it was further pruned excluding: structures containing cofactor molecules, more than one ligand, or metal ions in the binding site; proteins whose binding site is defined by more than one chain; multiple entries of the same protein. Moreover, complexes containing ligands with more than 10 rotatable bonds were excluded to reduce the CPU time needed for docking. The resulting reference set includes 21 structures of complexes, spanning several different protein folds and having different functions (see Table 4.1 for a description and Figure 4.1 and 4.2 for a view of the ligands and the protein structures).

Some of the proteins in the selected reference set are relevant drug targets: the neuraminidase from influenza virus (PDB ID: 1A4Q), the protein-tyrosine phosphatase 1B (1C83), that is a therapeutic target in several disease, including diabetes, cancer, and inflammation, the urokinase-type plasminogen activator (1EJN), whose inhibition represents a promising target for antimetastatic therapy, the acetyltransferase that catalyzes the O-acetylation of the antibiotic chloramphenicol (3CLA) and the progesterone and estrogen receptors (1A28 and 3ERT). The set is also characterized by ligands with broadly different chemical characteristics, from peptides to aromatic carboxylic acids, from carbohydrates to nucleobases. Among them, some are pharmaceutically relevant molecules, like for example 3,5,3’,5’-tetraiodo-L-thyronine

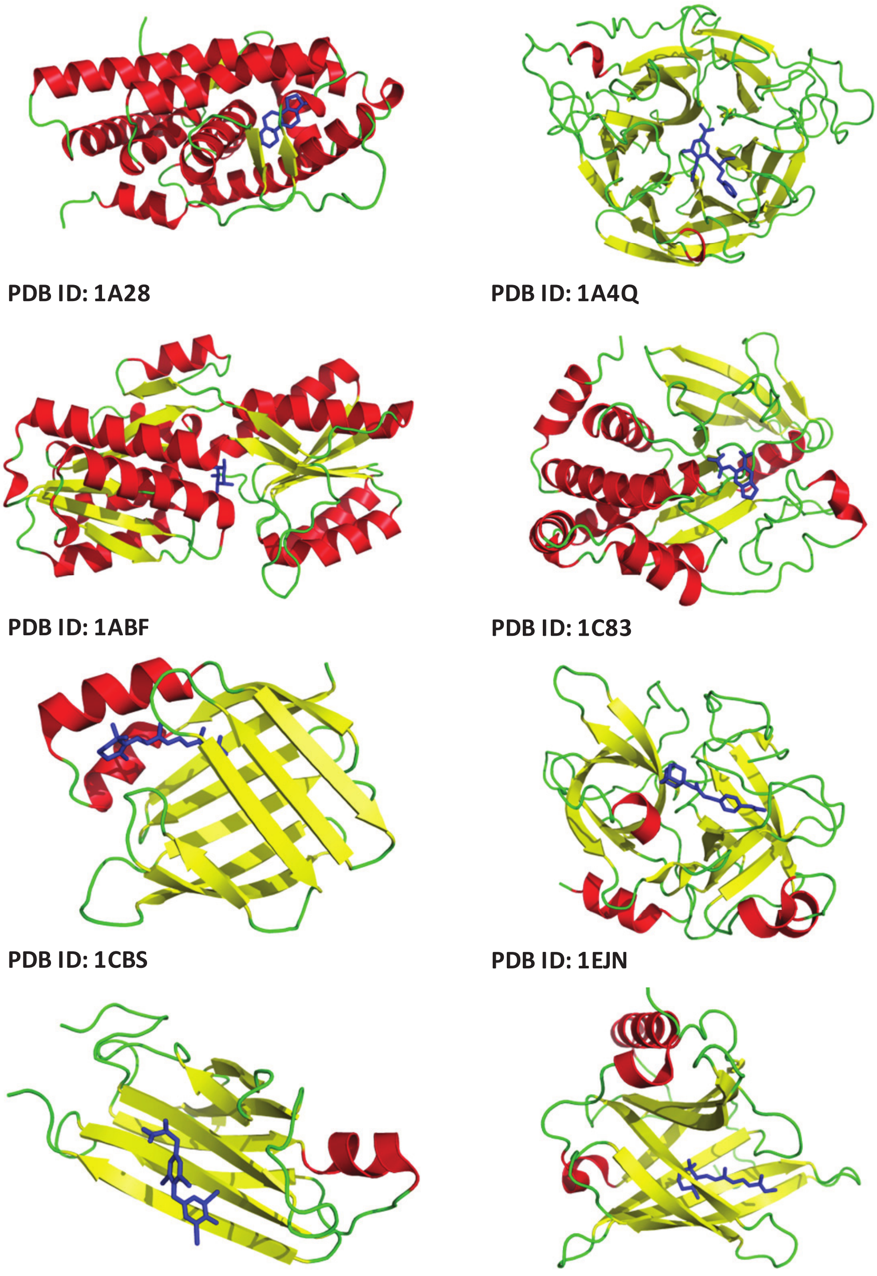
(PDB ligand identifier: T44), used to treat patients with thyroid problems, 4-hydroxytamoxifen (OHT), a selective estrogen receptor modulator, and chloramphenicol (CLM), a broad-spectrum antibiotic.

Table 4.1 – Test cases and selected templates. In bold are highlighted the templates for which an holo form was chosen for modelling. (*Continues*)

Complex PDB ID (chain ID)	Protein name	No. of residues	CATH ID	Ligand ID	Template PDB ID	Seq_Id (%)
1A28 (A)	Progesterone receptor	256	1.10.565.10	STR	**1NQ7**	17
					1YUC	21
					2AM9	55
1A4Q (A)	Neuraminidase	390	2.120.10.10	DPC	1V0Z	31
					2BAT	32
					2HT5	35
					2HTV	37
1ABF	L-arabinose binding protein	306	3.40.50.2300	FCA	**1TLF**	18
					2DRI	21
					2GBP	22
1C83 (A)	Protein-tyrosine phosphatase 1B	298	3.90.190.10	OAI	1RPM	34
					2GJT	37
1CBS	Cellular retinoic-acid-binding protein type II	137	2.40.128.20	REA	**1MDC**	24
					2FT9	27
					3IFB	30
					1FTP	38
					1CBI	77
1EJN (A)	Urokinase-type plasminogen activator	253	2.40.10.10	AGB	1OP8	23
					1PPF	26
					1ELT	33
					1YBW	38
1ETA (1)	Transthyretin (prealbumin)	127	2.60.40.180	T44	1OO2	56
					1TFP	79
					1IE4	84
1FEN	Retinol binding protein	183	2.40.128.20	AZE	1EXS	21
					1IIU	82
1LST	Lysine-, arginine-, ornithine-binding protein	239	3.40.109.10	LYS	**1XT8**	25
					1GGG	30
					1HSL	71
1MLD (A)	Malate dehydrogenase	314	3.40.50.720	CIT	**1I0Z**	19
					2LDX	23
					6LDH	24
					1HYG	27
					1SMK	56
					1EMD	59
1MRG	Alpha-momorcharin	263	3.40.420.10	ADN	1QI7	24
					1WUC	29
					1ABR	33
					1HWM	36
					1CF5	53
					1MRJ	64
					1BRY	66
					1NIO	70

Table 4.1 – (*Continued*)

Complex PDB ID (chain ID)	Protein name	No. of residues	CATH ID	Ligand ID	Template PDB ID	Seq_Id (%)
1MRK	Alpha-trichosanthin	247	3.40.420.10	FMC	1R4P	17
					1RL0	22
					1APA	29
					2MLL	33
					1RTC	37
					1NIO	60
					1CF5	62
					1MRG	64
1ROB	Ribonuclease A	124	3.10.130.10	C2P	**1OJ1**	27
					1DYT	28
					1ONC	30
					1B1I	32
					1AGI	36
					1RNF	42
					1Z7X	70
1SRJ (A)	Streptavidin	121	2.40.128.30	NAB	**1WBI**	27
					1Y52	30
1TNG	Trypsin	229	2.40.10.10	AMC	1QY6	15
					1A7S	30
					1FIW	34
					1YBW	38
					2F91	43
					1HJ8	66
					1H4W	73
					1A0J	74
					1TRN	75
					2A31	82
1UKZ	Uridylate kinase	203	3.40.50.300	AMP	**1G3U**	17
					1GKY	23
					1AKE	28
					1Z83	44
					1TEV	48
					1UKE	52
2AK3 (A)	Adenylate kinase isoenzyme-3	226	3.40.50.300	AMP	**1MV5**	18
					1VHL	19
					1TEV	25
					1Z83	26
					1UKE	27
					2AK2	40
					2AKY	41
					2AR7	58
3CLA	Chloramphenicol acetyltransferase	213	3.30.559.10	CLM	1NOC	47
3ERT (A)	Estrogen receptor alpha	261	1.10.565.10	OHT	**1NQ7**	21
					1PK5	24
6RNT	Ribonuclease T1	104	3.10.450.30	2AM	**1RMS**	65
7TIM (A)	Triosephosphate isomerase	247	3.20.20.70	PGH	1B9B	43
					1R2R	53
					1WYI	53
					1MO0	54



PDB ID: 1A28

PDB ID: 1A4Q

PDB ID: 1ABF

PDB ID: 1C83

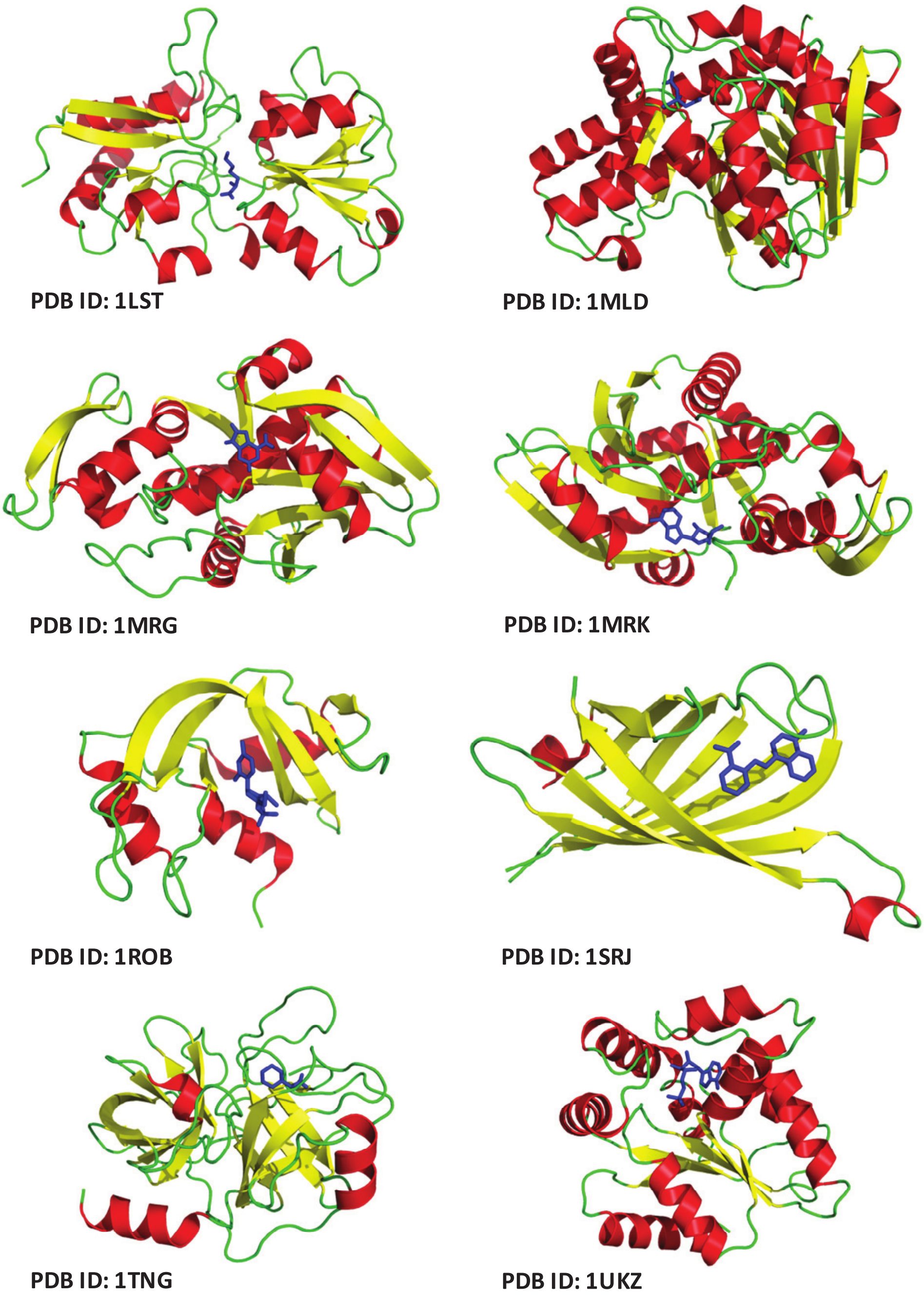
PDB ID: 1CBS

PDB ID: 1EJN

PDB ID: 1ETA

PDB ID: 1FEN

Figure 4.1 – 3D structures of the complexes in the test set. The proteins are represented in cartoon, the ligands in sticks (blue). The different secondary structure elements of the proteins are characterized by different colours: helices in red; strands in yellow. In green are represented the loop regions. The PDB ID of the complex is reported under each structure. (*Continues*)



PDB ID: 1LST

PDB ID: 1MLD

PDB ID: 1MRG

PDB ID: 1MRK

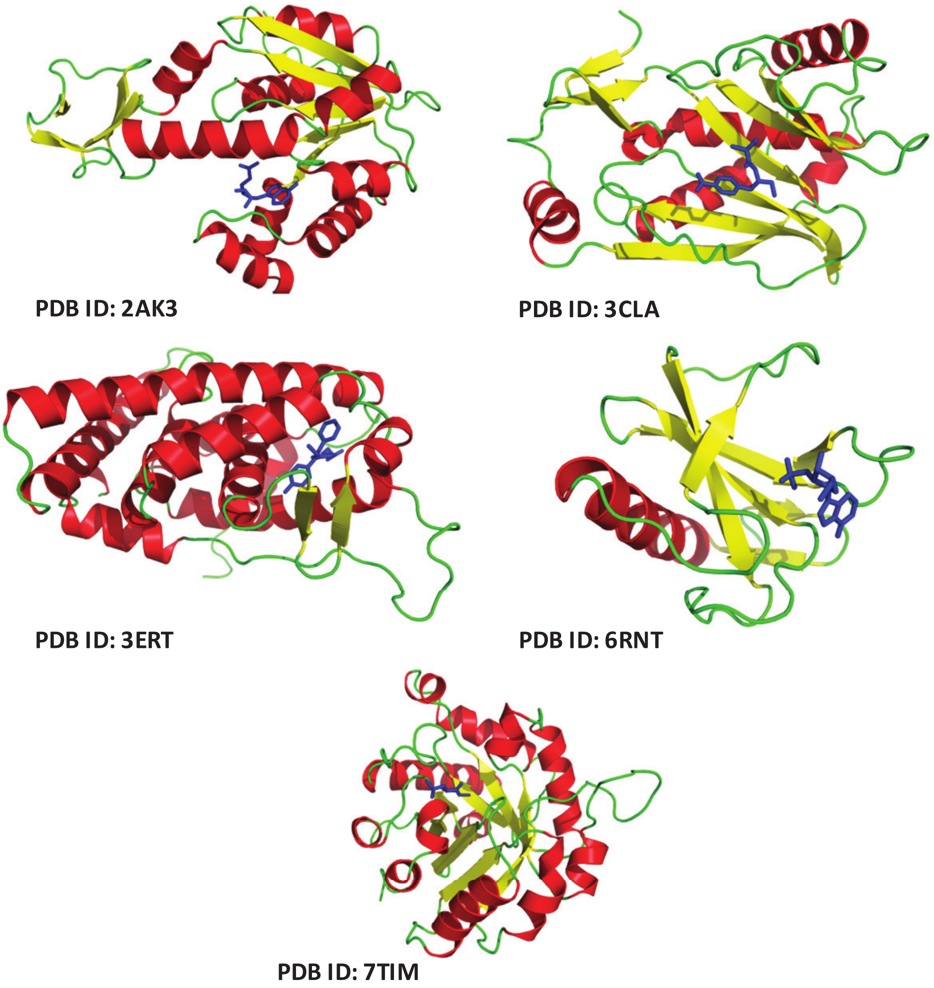
PDB ID: 1ROB

PDB ID: 1SRJ

PDB ID: 1TNG

PDB ID: 1UKZ

Figure 4.1 – (*Continued*)



PDB ID: 2AK3

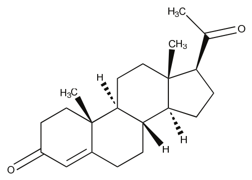
PDB ID: 3CLA

PDB ID: 3ERT

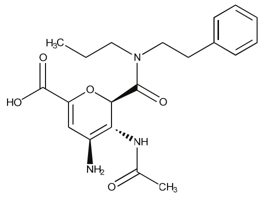
PDB ID: 6RNT

PDB ID: 7TIM

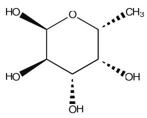
Figure 4.1 – (*Continued*)



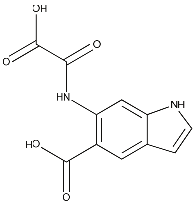
STR: PROGESTERONE



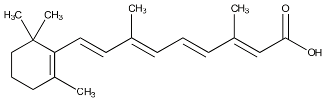
DPC: 5- ACETYLAMINO- 4- AMINO- 6- (PHENETHYL- PROPYL- CARBAMOYL)- 5,6- DIHYDRO- 4H- PYRAN- 2- CARBOXYLIC ACID



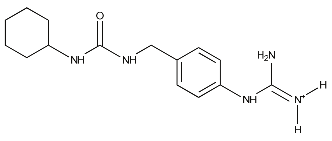
FCA: ALPHA- D- FUCOSE



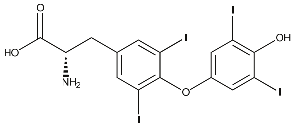
OAI: 6- (OXALYL- AMINO)- 1H- INDOLE- 5- CARBOXYLIC ACID



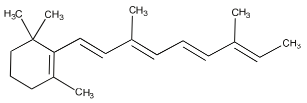
REA: RETINOIC ACID



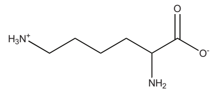
AGB: N- (1- ADAMANTYL)- N'- (4- GUANIDINOBENZYL) UREA



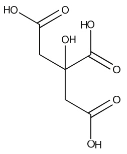
T44: 3,5,3',5'- TETRAIODO- L- THYRONINE



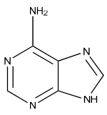
AZE: ALL- TRANS AXEROPHTHENE



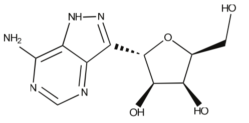
LYS: LYSINE



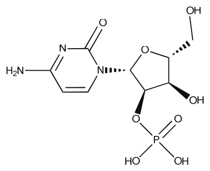
CIT: CITRIC ACID



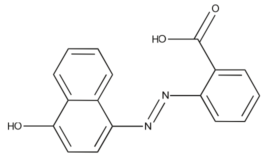
ADN: ADENOSINE



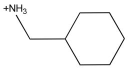
FMC: (1S)- 1- (7- AMINO- 1H- PYRAZOLO [4,3- D]PYRIMIDIN- 3- YL)- 1,4- ANHYDRO- D- RIBITOL



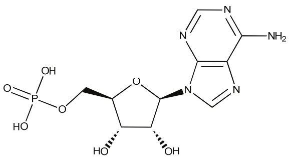
C2P: CYTIDINE- 2'- MONOPHOSPHATE



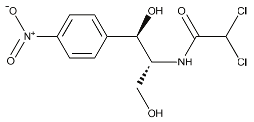
NAB: 2- ((4'- HYDROXYNAPHTHYL)- AZO) BENZOIC ACID

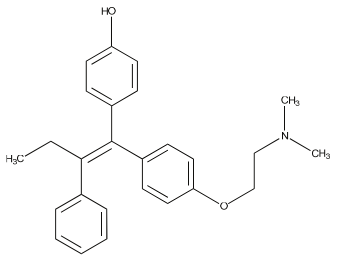


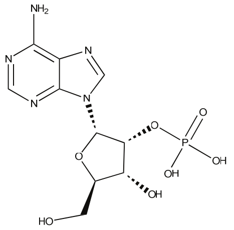
AMC: AMINOMETHYLCYCLOHEXANE



AMP: ADENOSINE MONOPHOSPHATE



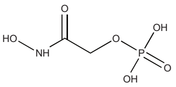




CLM: CHLORAMPHENICOL

OHT: 4- HYDROXYTAMOXIFEN

2AM: ADENOSINE- 2'- MONOPHOSPHATE



PGH: PHOSPHOGLYCOLO HYDROXAMIC ACID

Figure 4.2 – Chemical structures of the ligands in the test set. For each structure, the three-letter identifier and the chemical name associated to the ligand in the PDB are reported.

Homology modelling procedures

A schematic workflow of this work is reported in Figure 4.3. Two different strategies were employed to generate the modelling dataset: a fully automated modelling method, which employs a prediction server, and a traditional homology modelling procedure. This consists of three major steps: identification of candidate template structures, alignment of the target to the template and structural modelling of the target on the template structure.

The choice of the programs for each step and the details of the protocol for the traditional homology modelling procedure were based on recent assessments of homology model strategies [21] and template selection strategies [18], in order to reproduce standard homology modelling experiments as well as to obtain a large range of model quality.

Identification of candidate templates was performed by sequence similarity search using PSI-BLAST [142] with default parameters until convergence was reached. Each target was searched against a database of all proteins of known structure from the NCBI database. The resulting candidate lists were reduced by elimination of all hits having low statistical significance (BLAST E-value greater than 0.01) or alignment length shorter than 85% of the target sequence. A statistical analysis on the distribution of sequence identities between target and template was performed and the final subset of template candidates was selected to provide a reliable sampling of different evolutionary distances. In order to avoid backbone modelling errors in the binding site, all the selected templates do not contain gaps in the alignment of the active site region, this being defined as the list of residues with at least one atom within 4 Å of any of the ligand atoms. This step clearly reduced the candidate template list (e.g. 6RNT, where only one template remained), but it did not result in the complete exclusion of any of the original targets.

Alignments of target and template sequences were performed with three independent tools: a sequence-sequence (T-Coffee [143]), a profile-profile (PRALINE [144]) and a structure-structure (TM-align [80]) alignment method. They were selected in order to obtain sequence alignments between target and template at different levels of accuracy. T-Coffee was used for obtaining both single and multiple sequence alignments; this method carries out a progressive alignment driven by all the pair-wise local and global

sequence alignments. PRALINE, a dynamic programming-based method that employs a profile-based progressive sequence alignment protocol, was employed for multiple-template alignments. Finally, TM-align, a method to identify the best structural alignment between protein pairs that combines the TM-score rotation matrix and dynamic programming, was used for generating single-template structural alignments.
Model construction was performed by using MODELLER 9v1 [67], which implements an approach to comparative modelling by satisfying spatial restraints derived from the alignment of the target sequence with the template structure. The method is described in Paragraph 2.1.1.

For each target, an additional model was generated by using the automated server I-TASSER, that was ranked as the best method in the server section of the latest CASP experiments [145]. I-TASSER is a hierarchical protein structure modelling approach based on the secondary-structure enhanced Profile-Profile threading Alignment (PPA) and the iterative implementation of the Threading ASSEmbly Refinement (TASSER) program [70, 71] (for a detailed description, see Paragraph 2.1.2).

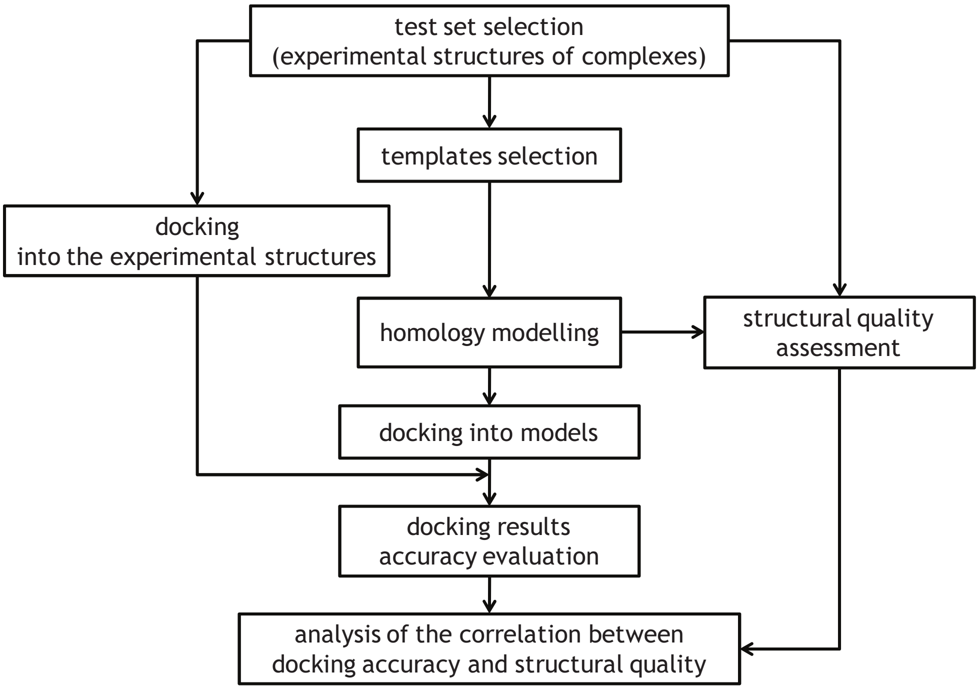

test set selection
(experimental structures of complexes)
templates selection
docking
into the experimental structures
homology modelling
structural quality
assessment
docking into models
docking results
accuracy evaluation
analysis of the correlation between
docking accuracy and structural quality

Figure 4.3 – Workflow of the project**.**

Model quality indices

Models were assessed both by direct comparison to the known native structures (‘calculated’ indices) and by using indices for model quality estimation and prediction (‘predicted’ indices).

Direct comparison of each model to the corresponding native structure was obtained by structural alignment using two programs: DALILite [83] and LGA [84]. The former generates a global alignment by optimization of both the number of structurally aligned residues and the overall root mean square deviation (RMSD) of their corresponding C-α atoms. The latter generates many different local superimpositions to detect regions where proteins are similar, providing a combination of two sets of scores for Local/Global Alignment (LGA). According to the structural alignment results, the global quality of the models was measured by six indices (see Paragraph 2.2): the *RMSD* on C-α atoms; *DALI_Zscore*; DALI_Zratio; *GDT_TS*; *GDT_HA*; *LGA_S*; *LGA_RMSD*.
The local quality was measured by three indices: the site RMSD based on DALI alignments (*RMSD-s*), the amino acids in the active site being defined as the residues with at least one atom within 4 Å of any of the ligand atoms; the dRMSD (distance Root Mean Square Deviation) among the heavy atoms of the sidechains in the active site (*dRMSD-s*); and the Atomic Contact Score (*ACS)* [87].

In the class of ‘predicted’ indices some indices that evaluate the model-template similarities were considered. On the basis of the T-Coffee alignments, the percentage of sequence identity and similarity for the whole sequence length (*Seq_Id, Seq_Sim*) and only for the active site (*Seq_Id-s, Seq_Sim-s*) were calculated for each target – template alignment. The following indices were also calculated for the model – template structural alignments generated with DALILite and LGA: *RMSD(t), RMSD-s(t), LGA_RMSD(t), LGA_S(t).*
Complementary to these indices is the set of ‘predicted’ indices derived by geometrical analysis of the models performed by submission to the Protein Structure Validation Software (PSVS) web server [89]. This server integrates analyses from several widely-used structure quality evaluation tools, including: PROCHECK [91], Verify3D [93], ProsaII [92], MolProbity [90]. From the extensive output of the server, the following indices were chosen for our analysis: Z scores from *ProsaII, Verify3D, MolProbity,* and the

Procheck(phi, psi) index (*Procheck*). Other methods among the recent proposals for Model Quality Assessment were also included in the analysis: ProQ [94], that predicts the final quality of a single model as the expected *LGscore* [95] and *MaxSub* [96] indices, and the ModFOLD server [97] that calculates the *MQscore* index.

A detailed description of the single indices can be found in Paragraph 2.2.

Molecular docking calculations

AutoDock 4 [29, 106] and its graphical front-end AutoDockTools (ADT) were used to set up and perform docking calculations (see Paragraph 2.3).

Experimental structures were downloaded from the Protein Data Bank while theoretical models were constructed as described above; for the former all water molecules, ions and ligands were removed from the original files. Polar hydrogen atoms were added to each protein structure (both crystallographic structures and models). The structures of the ligands were directly extracted from the CCDC/Astex Test Set [111], therefore their bond lengths and angles were already optimized; AutoDock tool AutoTors was employed to identify the ligands rotatable bonds. The proteins were treated as rigid bodies during docking simulations.

Grid maps with 0.375 Å spacing were defined for each collection of experimental structures and corresponding models to include the protein binding site. Lamarckian Genetic Algorithms, as implemented in AutoDock, were employed to perform docking calculations. The maximum number of energy evaluations and of generations were set to 25 million and to 27000 respectively and 100 runs, each with a population of 150 individuals, were performed for each calculation. Random starting positions, orientations and torsions were used for the ligands, their translational step was set to 0.2 Å, the quaternion step to 5.0˚ and the torsion step to 5.0˚. Cluster analysis was performed by AutoDock with a RMSD cut-off of 2 Å. All other parameters were default settings.

In order to evaluate the docking results, the dRMSD was calculated between the model ligand-site distances and the X-ray ligand-site corresponding distances. Using this index, the distance calculation takes into account only the deviation on the relative position of the ligand to the residues belonging to the binding site and not, as it is for the RMSD calculation, of the deviation on the absolute position of the ligand in the pose from the crystallographic one. Therefore, the structural differences between the model and the experimental structure are excluded from the evaluation of the quality of docking results.

For each docking case, the dRMSD was evaluated both for the lowest energy pose, according to the Autodock scoring function (*dRMSD*), and for the lowest dRMSD pose (*mindRMDS*), in order to investigate the correlation between the quality of the models and the accuracy of docking results.

4.1.2 Results

Test set variety

The reference group of 21 X-ray structures of ligand-protein complexes (Table 4.1) is composed by proteins having different chain lengths (100 to 400 residues) and a large spectrum of structural characteristics. Three different CATH classes (mainly α, mainly β, $\alpha-\beta$) [146] and a wide range of architectures and topologies are represented.

For each protein in the reference set, several theoretical models were developed by different homology modelling strategies. The template structures were selected to cover a wide range of evolutionary distances with the target. A preliminary analysis of the distribution of the sequence identities of the candidate templates highlighted a bimodal trend for each similarity search result, with the highest peak at 20-30% identity and the second one at high values (50-80%) (one example is reported in Fig. 4.4a). In order to reproduce a similar distribution in the test case, templates were selected in a range of identity from 15 to 85%, accurately enriching the number of representatives for low percentages (see Fig. 4.4b), since these are the most difficult cases, on which it would be interesting to know if docking can give accurate results. The 92 selected templates (Table 4.1) include both apo and holo structures, the latter characterized by ligands which are usually different from the ones bound in reference structures.

Model quality variety

For each target several models were generated using Modeller: two single-template models for each selected template and two multiple-template models; moreover, an additional model was obtained using the automated server I-TASSER. The resulting set includes 245 models.

The quality of each model was at first evaluated by direct comparison with the native structure, employing the ‘calculated’ indices reported in Paragraph 4.1.1.

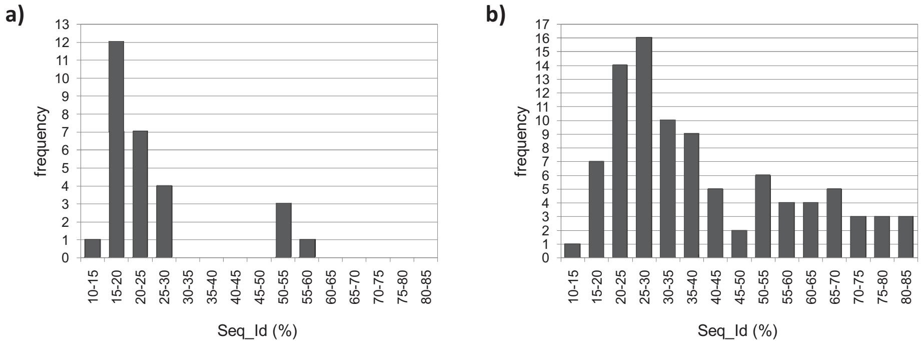


Figure 4.4 – a) Sequence identity distribution of the templates found for one of the test set cases (Progesterone receptor – reference PDB ID: 1A28), reported here as an example. b) Sequence identity distribution of the templates chosen for the project (see text).

The modelled set provides a wide spectrum of both global and binding site quality. Some examples are shown in Figure 4.5. The collection of modelled structures includes a significant number of good predictions and medium-low resolution models: the majority of *RMSD* values (Fig. 4.5a) are below 2 Å, while only 30% of the models show a lower conformity to the target, with *RMSD* values in the range 2-4 Å, and a few exceptions are of very low quality. Accordingly, with a small exception of 4 models, the *DALI_Zscore* values are always higher than 10, demonstrating that the models have the correct fold with statistical significance [83]. All the *DALI_Zratio* values are instead lower than 1, indicating a better agreement of the model to the template than to the native structure for all the cases. The trends of the three global indices based on LGA alignment are similar to those shown for DALI indices: the histogram of the *GDT_HA* values (Fig. 4.5b) shows that the majority of models have scores over 50, indicating a good similarity to the target, while 30% are in the range 20-50, and few models show poor conformity to the native structure.

This picture is slightly different for the active site where the distribution of site RMSD values (*RMSD-s,* Fig. 4.5c) is slightly skewed compared to the one for global *RMSD* and the group of models with a poorly reproduced binding site (about 10%, with values greater than 4 Å) does not include all the models with the worst *RMSD*. On the other hand, the *ACS* index of binding site quality shows a more uniform distribution (Fig. 4.5d), with about 40% of cases having binding sites modelled with high accuracy (*ACS* > 0.6), about 40%

with intermediate quality and the remaining cases with incorrectly modelled binding sites (*ACS* < 0.2).

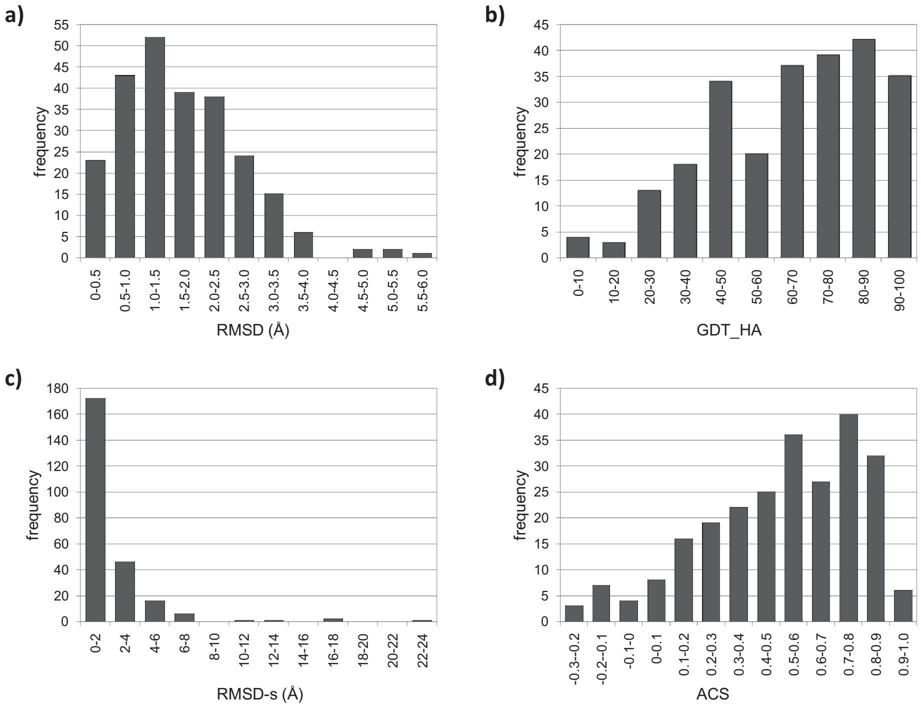


Figure 4.5 – Distributions of some 'calculated' indices of model quality for the modelling set: a) *RMSD*, b) *GDT_HA*, c) *RMSD-s*, d) *ACS*.

Correlations between 'calculated' indices

The degree of correlation between the 'calculated' indices was measured by Spearman's rank coefficient, ρ and the absolute values are reported in Table 4.2. This coefficient is generally more appropriate than Pearson's for handling non-normal distributions, as those we obtained for some indices (see for example the *RMSD-s* in Figure 4.5c). Values in Table 4.2 show that the indices calculated on the basis of the LGA structural alignment are highly correlated to each other (ρ absolute values from 0.97 to 1.00), as the elements of this group provide a very similar ranking of global model quality. Interestingly, *DALI_Zratio* is also highly correlated to both the *GDT_TS* and the *GDT_HA* scores (ρ = 0.90). On the basis of this correlation analysis, only the results for a representative index of this group, *GDT_HA,* will be reported and discussed in the following. On the contrary, all the three indices related to the binding site quality will be considered, since each of them seems to provide an information partially different from the others.

Lower ρ absolute values are observed when comparing indices of global quality with those of site quality. This confirms that there is no general trend for models with accurate backbone geometry to have accurately modelled binding sites [87].

Correlations between 'predicted' and 'calculated' indices

The analysis of the pairwise correlations between the 'predicted' indices indicated in Paragraph 4.1.1 (Table 4.2) shows high correlation values (|ρ| > 0.90) between *Seq_Id* and *Seq_Sim, Seq_Id-s* and *Seq_Sim-s*, as well as between model-template RMSD calculated by using DALI (*RMSD*(t)) or obtained from the LGA alignment (*LGA_RMSD(t)*). The two indices calculated with the model quality assessment program ProQ (*LGscore* and *MaxSub*) were also highly correlated.

The identification of the most effective indices for quality prediction was based on the Spearman correlation coefficients for pairwise comparisons of 'predicted' and 'calculated' indices (Table 4.2): the 'predicted' indices with the highest correlation with a 'calculated' index are expected to be more powerful in prediction. In particular, the indices having |ρ| > 0.70 with respect to both *RMSD* and *GDT_HA*, taken as a reference, are: *Seq_Id* (and the correlated *Seq_Sim*), *RMSD(t)* (and *LGA_RMSD(t)*), the corresponding indices evaluated in the binding site, the *MolProbity* and *Verify3D* Z-scores. The expected relationship between target-template sequence identity and the *RMSD* [17] is observed (Fig. 4.6a), and the plot confirms that, also in this set of structures, models with *Seq_Id* greater than 50% have *RMSD* values lower than 2 Å. On the other hand, *Seq_Id* shows lower correlation coefficients with respect to the 'calculated' site quality indices, *RMSD-s, dRMSD-s* and *ACS* (Table 4.2). The plot in Figure 4.6b highlights that models with *Seq_Id* values both greater and lower than 50% can give accurately modelled binding sites.

The same analysis described above for the whole test set was performed on the two subsets of models generated starting from T-Coffee and TM-align pair-wise sequence alignments. A wide spectrum of model quality is obtained also from models generated using a single strategy. Moreover, the expected relationship between the target-template sequence identity and the *RMSD* was also found for these subsets of models and the trend of relation between sequence identity and *RMSD-s* showed by the whole set of models was present also in these cases (data not reported).

Table 4.2 – Correlations (absolute values of Spearman's rank coefficient) between model quality indices for the whole test set. The darker colours represent stronger correlations.

	RMSD	DALI_Zscore	DALI_Zratio	LGA_RMSD	GDT_TS	GDT_HA	LGA_S	RMSD-s	dRMSD-s	ACS	Seq_Id	Seq_Sim	RMSD(t)	LGA_RMSD(t)	LGA_S(t)	Seq_Id-s	Seq_Sim-s	RMSD-s(t)	MolProbity	Procheck	Prosall	Verify3D	LGscore	MaxSub	MQscore
RMSD	1.00																								
DALI_Zscore	0.63	1.00																							
DALI_Zratio	0.88	0.63	1.00																						
LGA_RMSD	0.87	0.64	0.86	1.00																					
GDT_TS	0.89	0.67	0.90	0.98	1.00																				
GDT_HA	0.89	0.66	0.90	0.98	1.00	1.00																			
LGA_S	0.88	0.66	0.89	0.97	0.99	0.99	1.00																		
RMSD-s	0.87	0.59	0.76	0.81	0.82	0.82	0.81	1.00																	
dRMSD-s	0.73	0.40	0.57	0.66	0.66	0.67	0.65	0.84	1.00																
ACS	0.75	0.51	0.67	0.68	0.69	0.71	0.68	0.83	0.78	1.00															
Seq_Id	0.85	0.54	0.69	0.86	0.86	0.87	0.85	0.73	0.68	0.69	1.00														
Seq_Sim	0.84	0.51	0.65	0.84	0.85	0.86	0.84	0.70	0.65	0.66	0.98	1.00													
RMSD(t)	0.76	0.63	0.60	0.76	0.77	0.78	0.76	0.61	0.51	0.52	0.86	0.86	1.00												
LGA_RMSD(t)	0.77	0.65	0.62	0.78	0.79	0.79	0.78	0.61	0.53	0.54	0.87	0.88	0.97	1.00											
LGA_S(t)	0.26	0.12	0.20	0.29	0.30	0.32	0.28	0.24	0.20	0.24	0.41	0.41	0.42	0.41	1.00										
Seq_Id-s	0.76	0.59	0.65	0.77	0.78	0.79	0.76	0.69	0.71	0.71	0.85	0.81	0.71	0.73	0.20	1.00									
Seq_Sim-s	0.69	0.59	0.57	0.71	0.72	0.72	0.71	0.65	0.66	0.67	0.81	0.78	0.70	0.71	0.19	0.90	1.00								
RMSD-s(t)	0.72	0.57	0.60	0.69	0.71	0.71	0.70	0.69	0.66	0.65	0.74	0.73	0.82	0.82	0.31	0.70	0.67	1.00							
MolProbity	0.81	0.29	0.69	0.75	0.76	0.78	0.76	0.71	0.63	0.64	0.84	0.84	0.62	0.63	0.33	0.69	0.62	0.59	1.00						
Procheck	0.06	0.13	0.03	0.12	0.11	0.10	0.13	0.04	0.04	0.08	0.09	0.14	0.14	0.15	0.05	0.01	0.10	0.13	0.14	1.00					
Prosall	0.55	0.28	0.53	0.60	0.60	0.60	0.60	0.55	0.53	0.37	0.57	0.58	0.53	0.53	0.22	0.52	0.49	0.50	0.58	0.45	1.00				
Verify3D	0.76	0.67	0.72	0.72	0.74	0.74	0.73	0.64	0.55	0.55	0.76	0.74	0.74	0.76	0.22	0.68	0.61	0.63	0.63	0.05	0.52	1.00			
LGscore	0.54	0.84	0.54	0.57	0.59	0.58	0.60	0.50	0.35	0.40	0.51	0.52	0.61	0.64	0.13	0.48	0.50	0.55	0.36	0.42	0.49	0.66	1.00		
MaxSub	0.59	0.70	0.59	0.63	0.66	0.65	0.66	0.52	0.38	0.43	0.59	0.60	0.61	0.63	0.18	0.51	0.55	0.55	0.46	0.48	0.58	0.63	0.89	1.00	
MQscore	0.56	0.52	0.58	0.56	0.60	0.59	0.61	0.48	0.45	0.34	0.60	0.62	0.68	0.70	0.29	0.53	0.53	0.61	0.52	0.39	0.70	0.64	0.65	0.67	1.00

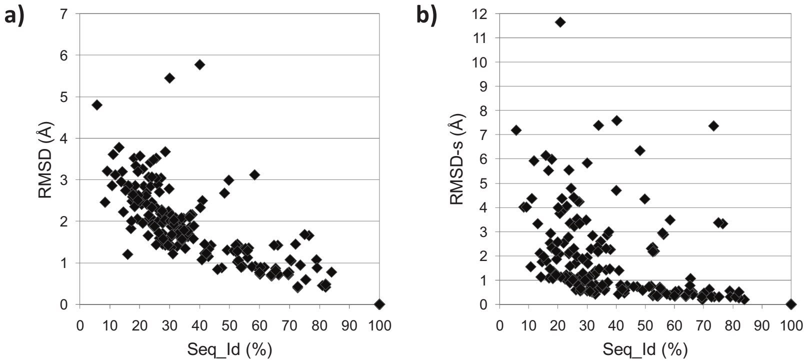


Figure 4.6 – Correlation between some 'calculated' model quality indices (a) *RMSD*, b) *RMSD-s*) and the model-template sequence identity (*Seq_Id*). Four cases with *RMSD-s* values higher than 12 Å were excluded from the plot b) for a clearer representation.

The results of the analysis of the correlation between all the quality indices considered in this work are reported in Table 4.3 and 4.4 for the T-Coffee and the TM-align subsets, respectively. The same observation as for the whole set of data can be made for these subsets. The main difference between the subsets and the whole set (Fig. 4.7 and Tables 4.2, 4.3 and 4.4) is the higher correlation found between the *MQscore* and the LGA indices (*LGA_RMSD, GDT_TS, GDT_HA and LGA_S*) for the T-Coffee subset.

Docking results accuracy

The docking experiments were aimed at reproducing the binding geometries corresponding to all the protein-ligand complexes in the reference set. To this end, molecular docking calculations were performed on both the protein experimental structures and the associated group of structural models, for a total of 266 simulations. During the docking process, the protein was treated as a rigid body and only the ligand flexibility was considered, by including the torsional degrees of freedom of all the ligand rotatable bonds. Cluster analysis of the poses was performed using AutoDock for each docking simulation.

In the majority of the cases, the results indicated an efficient sampling of the inter- and intra-molecular conformational space, associated with a reduced number of highly populated clusters of poses. In some cases (around 20% of the total docking runs), the first cluster was scarcely populated, thus indicating a poor sampling convergence.

Table 4.3 – Correlations (absolute values of Spearman's rank coefficient) between model quality indices for the models generated from T-Coffee alignments (see text). The darker colours represent stronger correlations.

	RMSD	DALI_Zscore	DALI_Zratio	LGA_RMSD	GDT_TS	GDT_HA	LGA_S	RMSD-s	dRMSD-s	ACS	Seq_Id	Seq_Sim	RMSD(t)	LGA_RMSD(t)	LGA_S(t)	Seq_Id-s	Seq_Sim-s	RMSD-s(t)	MolProbity	Procheck	Prosall	Verify3D	LGscore	MaxSub	MQscore
RMSD	1.00																								
DALI_Zscore	0.66	1.00																							
DALI_Zratio	0.94	0.74	1.00																						
LGA_RMSD	0.82	0.65	0.86	1.00																					
GDT_TS	0.87	0.69	0.91	0.97	1.00																				
GDT_HA	0.87	0.69	0.92	0.97	1.00	1.00																			
LGA_S	0.86	0.69	0.90	0.96	0.99	0.99	1.00																		
RMSD-s	0.77	0.51	0.75	0.75	0.76	0.76	0.75	1.00																	
dRMSD-s	0.66	0.36	0.64	0.64	0.65	0.65	0.65	0.82	1.00																
ACS	0.72	0.53	0.73	0.63	0.66	0.67	0.65	0.77	0.77	1.00															
Seq_Id	0.79	0.59	0.82	0.79	0.84	0.85	0.84	0.64	0.64	0.60	1.00														
Seq_Sim	0.76	0.55	0.79	0.78	0.84	0.84	0.84	0.61	0.61	0.55	0.97	1.00													
RMSD(t)	0.74	0.62	0.70	0.75	0.77	0.78	0.78	0.57	0.53	0.49	0.85	0.85	1.00												
LGA_RMSD(t)	0.73	0.62	0.70	0.75	0.77	0.78	0.78	0.56	0.53	0.49	0.86	0.86	0.98	1.00											
LGA_S(t)	0.32	0.16	0.33	0.41	0.43	0.45	0.44	0.29	0.29	0.26	0.50	0.49	0.51	0.53	1.00										
Seq_Id-s	0.70	0.62	0.75	0.70	0.74	0.75	0.73	0.62	0.64	0.64	0.84	0.78	0.69	0.69	0.28	1.00									
Seq_Sim-s	0.63	0.62	0.69	0.66	0.69	0.69	0.69	0.58	0.59	0.61	0.81	0.77	0.67	0.67	0.24	0.88	1.00								
RMSD-s(t)	0.75	0.55	0.71	0.69	0.73	0.75	0.74	0.65	0.69	0.66	0.77	0.77	0.84	0.84	0.39	0.69	0.64	1.00							
MolProbity	0.71	0.34	0.72	0.66	0.72	0.73	0.72	0.55	0.56	0.53	0.84	0.85	0.70	0.69	0.42	0.66	0.64	0.70	1.00						
Procheck	0.04	0.11	0.05	0.18	0.15	0.14	0.18	0.04	0.09	0.05	0.15	0.21	0.20	0.20	0.05	0.04	0.15	0.18	0.23	1.00					
Prosall	0.55	0.30	0.57	0.74	0.72	0.72	0.72	0.58	0.60	0.42	0.66	0.67	0.60	0.59	0.34	0.56	0.51	0.58	0.65	0.43	1.00				
Verify3D	0.75	0.73	0.79	0.71	0.76	0.77	0.75	0.55	0.52	0.52	0.75	0.73	0.77	0.76	0.31	0.67	0.61	0.70	0.61	0.06	0.56	1.00			
LGscore	0.60	0.83	0.66	0.65	0.67	0.66	0.68	0.44	0.34	0.40	0.57	0.57	0.62	0.62	0.18	0.51	0.52	0.55	0.40	0.43	0.52	0.72	1.00		
MaxSub	0.61	0.69	0.67	0.67	0.71	0.70	0.71	0.46	0.38	0.41	0.64	0.64	0.63	0.63	0.22	0.53	0.56	0.56	0.48	0.42	0.58	0.67	0.85	1.00	
MQscore	0.60	0.57	0.65	0.74	0.76	0.75	0.78	0.49	0.54	0.42	0.76	0.77	0.76	0.77	0.44	0.61	0.62	0.67	0.63	0.45	0.74	0.74	0.75	0.75	1.00

Table 4.4 – Correlations (absolute values of Spearman's rank coefficient) between model quality indices for the models generated from TM-align alignments (see text). The darker colours represent stronger correlations**.**

	RMSD	*DALI_Zscore*	*DALI_Zratio*	*LGA_RMSD*	*GDT_TS*	*GDT_HA*	*LGA_S*	*RMSD-s*	*dRMSD-s*	*ACS*	*Seq_Id*	*Seq_Sim*	*RMSD(t)*	*LGA_RMSD(t)*	*LGA_S(t)*	*Seq_Id-s*	*Seq_Sim-s*	*RMSD-s(t)*	*MolProbity*	*Procheck*	*Prosall*	*Verify3D*	*LGscore*	*MaxSub*	*MQscore*
RMSD	1.00																								
DALI_Zscore	0.60	1.00																							
DALI_Zratio	0.88	0.49	1.00																						
LGA_RMSD	0.92	0.62	0.86	1.00																					
GDT_TS	0.94	0.63	0.87	0.98	1.00																				
GDT_HA	0.94	0.61	0.88	0.98	0.99	1.00																			
LGA_S	0.93	0.64	0.86	0.96	0.99	0.97	1.00																		
RMSD-s	0.79	0.58	0.76	0.81	0.81	0.81	0.79	1.00																	
dRMSD-s	0.57	0.33	0.49	0.61	0.58	0.60	0.56	0.77	1.00																
ACS	0.63	0.46	0.64	0.69	0.68	0.70	0.66	0.85	0.72	1.00															
Seq_Id	0.81	0.54	0.67	0.88	0.87	0.89	0.85	0.68	0.56	0.61	1.00														
Seq_Sim	0.81	0.52	0.63	0.87	0.86	0.87	0.85	0.66	0.52	0.59	0.98	1.00													
RMSD(t)	0.79	0.68	0.62	0.86	0.87	0.86	0.86	0.72	0.55	0.58	0.86	0.87	1.00												
LGA_RMSD(t)	0.80	0.72	0.61	0.86	0.87	0.86	0.86	0.72	0.56	0.61	0.87	0.88	0.95	1.00											
LGA_S(t)	0.21	0.14	0.22	0.27	0.28	0.28	0.26	0.27	0.20	0.25	0.30	0.30	0.29	0.26	1.00										
Seq_Id-s	0.68	0.58	0.58	0.77	0.75	0.77	0.72	0.64	0.65	0.65	0.81	0.77	0.74	0.76	0.13	1.00									
Seq_Sim-s	0.63	0.58	0.49	0.70	0.68	0.69	0.66	0.61	0.62	0.63	0.76	0.74	0.73	0.74	0.11	0.89	1.00								
RMSD-s(t)	0.69	0.62	0.53	0.72	0.71	0.71	0.70	0.75	0.64	0.65	0.69	0.69	0.80	0.80	0.23	0.71	0.68	1.00							
MolProbity	0.70	0.13	0.66	0.73	0.73	0.75	0.72	0.57	0.45	0.53	0.79	0.78	0.58	0.59	0.33	0.59	0.49	0.49	1.00						
Procheck	0.07	0.08	0.05	0.09	0.05	0.03	0.11	0.01	0.01	0.06	0.13	0.19	0.08	0.12	0.06	0.02	0.11	0.09	0.12	1.00					
Prosall	0.54	0.20	0.50	0.56	0.56	0.55	0.59	0.43	0.42	0.34	0.53	0.56	0.51	0.49	0.17	0.49	0.49	0.42	0.55	0.42	1.00				
Verify3D	0.76	0.63	0.67	0.79	0.81	0.80	0.80	0.57	0.41	0.47	0.75	0.74	0.76	0.78	0.19	0.62	0.53	0.57	0.55	0.00	0.41	1.00			
LGscore	0.53	0.85	0.39	0.55	0.57	0.54	0.60	0.47	0.21	0.37	0.50	0.52	0.65	0.69	0.15	0.45	0.48	0.58	0.15	0.32	0.39	0.61	1.00		
MaxSub	0.59	0.71	0.45	0.62	0.63	0.61	0.67	0.48	0.25	0.43	0.61	0.65	0.66	0.69	0.23	0.49	0.56	0.55	0.35	0.46	0.54	0.61	0.90	1.00	
MQscore	0.61	0.50	0.46	0.61	0.64	0.63	0.67	0.45	0.37	0.31	0.62	0.65	0.70	0.71	0.21	0.50	0.51	0.59	0.41	0.30	0.66	0.58	0.60	0.62	1.00

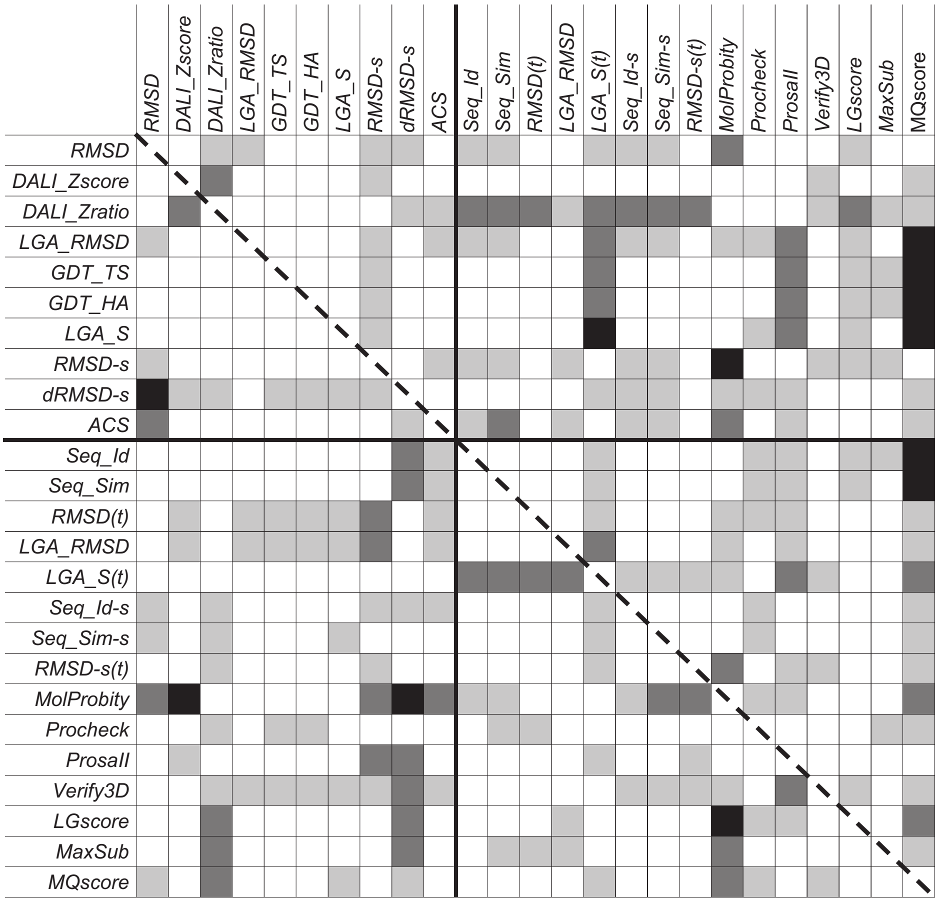


Figure 4.7 – Differences in correlations within model quality indices (see text). In the upper half of the matrix, as divided by the dashed line, difference between the absolute values of ρ Spearman obtained for the whole modelled set and for the T-Coffee models. In the lower half of the matrix, difference between the absolute values of ρ Spearman obtained for the whole modelled set and for the TM-align models (lower half of the matrix). The colours represent different degree of differences: in white |Δρ| < 0.05; in light grey 0.05 ≤ |Δρ| < 0.10; in dark grey 0.10 ≤ |Δρ| < 0.15; in black |Δρ| ≥ 0.15.

The accuracy of the geometries obtained was evaluated by calculating the dRMSD between the model ligand-site distances and the X-ray ligand-site corresponding distances. The distribution of the *dRMSD* values for the best-scored docking poses is reported in Figure 4.8. In more than 50% of the experiments the top ranked docking pose reproduces the experimental geometry with good accuracy (*dRMSD* < 3 Å). About 25% of the *dRMSD* values are in the range 3- 5 Å, whereas in the remaining cases high values (dRMSD > 5 Å) are observed. As expected, this last group includes the entries for which a poor convergence of the sampling was observed. As an example of the values of dRMSD associated to different outcomes of docking calculations, the binding geometries obtained in four docking simulations of the same complex (PDB ID: 1UKZ), are shown in Figure 4.9. A satisfactory reproduction of the experimental geometry is associated with *dRMSD* < 3 Å and an accurate description of the backbone of the native binding site in the model (Fig. 4.9a and 4.9b). *dRMSD* values in the intermediate range 3-5 Å (Fig. 4.9c) indicate a slightly misplaced location of the ligand in the binding site, associated with translational and/or rotational displacements from the experimental binding geometry and a minor displacement of the backbone of the native binding site in the model. For *dRMSD* > 5 Å (Fig. 4.9d) docking failed to reproduce the binding geometry and a completely incorrect backbone of the native binding site is present in the modelled structure.

For three complexes (PDB ID: 1C83, 3CLA, 6RNT), binding geometries very different from the experimental ones were obtained even when ligands were docked into the protein X-ray structure, obtaining RMSD values > 2 Å for both the best-scored and the most adherent to the experimental binding geometry poses. Accordingly, the poses obtained for the associated modelled structures were incorrect, too. The results obtained for these three complexes (18 cases) were excluded from the analysis on the relationships between model quality and docking results accuracy, because the performance associated with these results would be unrelated to the quality of the modelled structure.

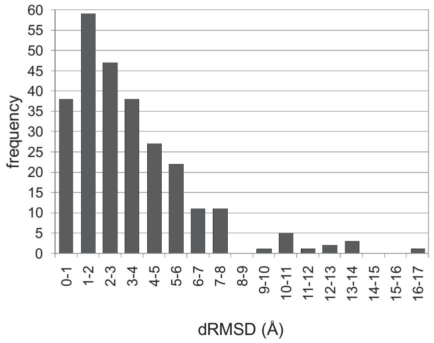


Figure 4.8 – Distribution of the *dRMSD* index of docking results for the whole test set.

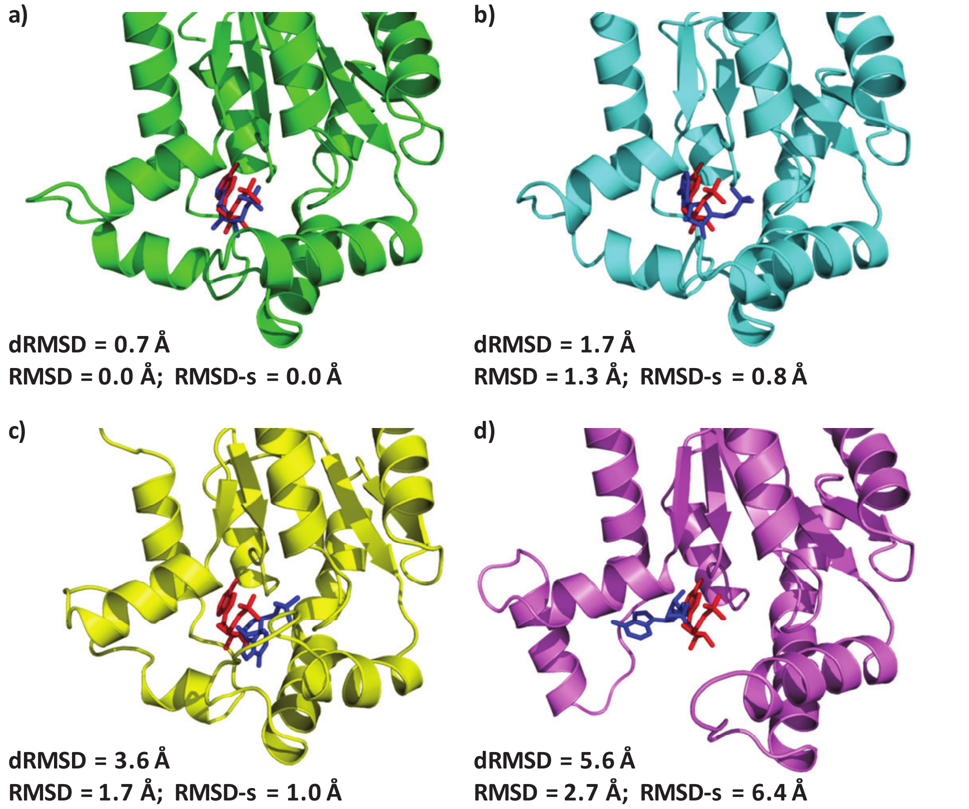

a)
dRMSD = 0.7 Å
RMSD = 0.0 Å; RMSD-s = 0.0 Å
b)
dRMSD = 1.7 Å
RMSD = 1.3 Å; RMSD-s = 0.8 Å
c)
dRMSD = 3.6 Å
RMSD = 1.7 Å; RMSD-s = 1.0 Å
d)
dRMSD = 5.6 Å
RMSD = 2.7 Å; RMSD-s = 6.4 Å

Figure 4.9 – Uridylate kinase – adenosine monophosphate complex (PDB ID: 1UKZ). Binding geometries obtained by docking calculations (blue sticks) compared to the experimental geometry (red sticks): a) ligand docking pose for the protein experimental structure (green cartoons); b) ligand docking pose for a protein model of high quality (cyan cartoons); c) ligand docking pose for a protein model of medium quality (yellow cartoons); d) ligand docking pose for a protein model of low quality (violet cartoons). Under each structure are reported the accuracy of docking results (*dRMSD*) and two 'calculated' indices of model quality: *RMSD* and *RMSD-s*.

It is conceivable that some observed inefficiencies of docking calculations are associated with the well known general limitation of the scoring functions implemented in docking programs: while the sampling procedure is able to generate good results, the scoring scheme is currently able to discriminate active from non-active in an ensemble of ligands, but it is often incorrect in ranking the binding poses generated for the same ligand and, consequently, in discriminating the 'true' pose from the others [34, 36]. The hypothesis that this limitation could affect the results of this work was investigated by analysing the relation between the dRMSD values for the best-scored pose (*dRMSD*) and the absolute minimum dRMSD to the crystallographic geometry that was obtained in our docking runs (*mindRMSD*). The plot in Figure 4.10a confirms that, in many cases, the best-scored pose does not correspond to the best geometrical pose and highlights that this trend is observed in the whole range of values. On the other hand, Figure 4.10b shows that the deviation between the scores of the first ranked poses and those of the best geometrical poses is generally limited.

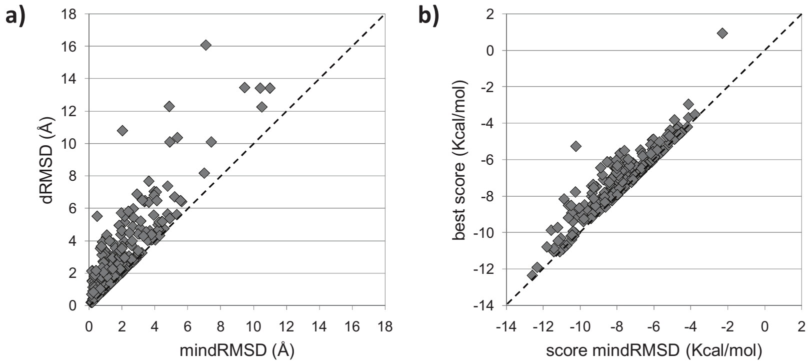


Figure 4.10 – Docking results: a) relation between dRMSD values for the best scored poses (*dRMSD*) and the minimum dRMSD values obtained (*mindRMSD*); b) relation between the AutoDock scores for the best scored poses (*best score*) and AutoDock scores for poses with the minimum dRMSD values (*score mindRMSD*).

Relationships between model quality and docking results accuracy

In addition to the test set, three structurally homogeneous subsets of complexes were analyzed independently to provide an insight on the role of fold specificity. Each subset was assembled selecting complexes whose proteins share the same fold and are evolutionary related, as indicated by the CATH classification in the same Homologous Superfamily (see Table 4.1):

- 2AK3 + 1UKZ: Adenylate kinase isoenzyme-3 – Adenosine monophosphate complex + Uridylate kinase – Adenosine monophosphate complex [CATH ID: 3.40.50.300];
- 1CBS + 1FEN: Cellular retinoic-acid-binding protein type II – Retinoic acid complex + Retinol binding protein - All-trans axerophthene complex [CATH ID: 2.40.128.20];

- 1EJN + 1TNG: Urokinase-type plasminogen activator – N-(1-adamantyl)-N'-(4-guanidinobenzyl) urea complex + Trypsin – Aminomethylcyclohexane complex [CATH ID: 2.40.10.10].

The results of the correlation analysis of the ‘calculated’ and ‘predicted’ indices with docking accuracy are reported in Table 4.5. The correlations were calculated with the *dRMSD* of the best-scored docking pose and, in addition, with the dRMSD of the poses that reproduced at best the experimental binding geometry (*mindRMSD*).

In the analysis of the whole test set, satisfactory correlations were obtained between *dRMSD* and all the ‘calculated’ indices of model quality, with the highest values for the site quality indices. The plots of the docking *dRMSD vs*. the *GDT_HA, RMSD-s* and *dRMSD-s* indices are shown as examples in Figure 4.11. As it is shown from data in Table 4.5, correlations are higher for the subsets of proteins belonging to the same fold than for the entire test set or the T-Coffee and TM-align subsets, both for the site and the global quality indices.

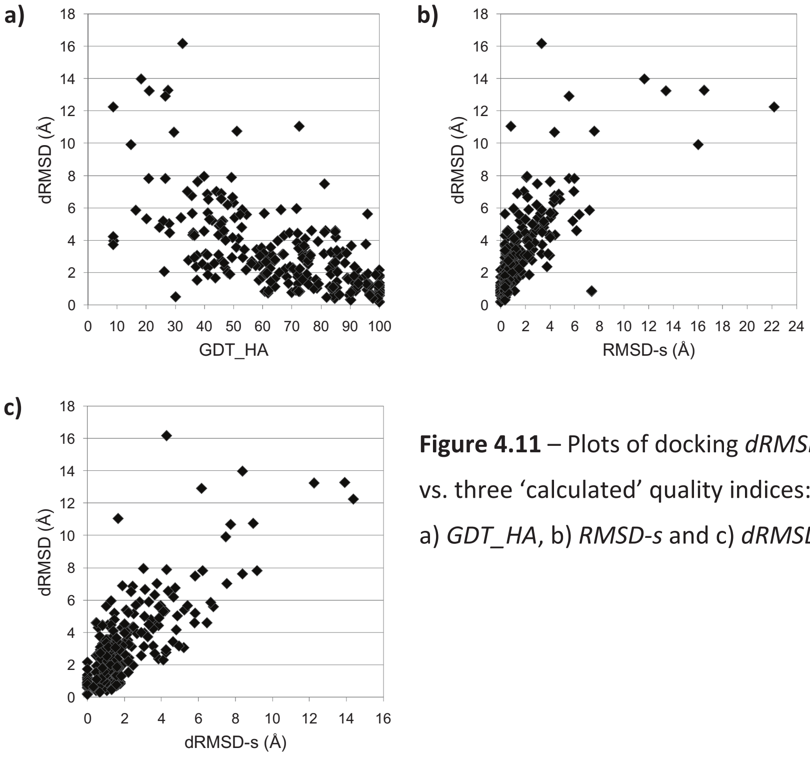


Figure 4.11 – Plots of docking *dRMSD* vs. three ‘calculated’ quality indices: a) *GDT_HA*, b) *RMSD-s* and c) *dRMSD-s*.

Table 4.5 – Correlations (absolute values of Spearman's rank coefficient) between some selected model quality indices and docking *dRMSD* and *mindRMSD* (in brackets), for the whole test set and for some subsets (see text).

	Whole test set (248 cases)	**T-Coffee set (89 cases**	**TM-align set (89 cases)**	**2AK3+1UKZ (36 cases)**	**1CBS+1FEN (22 cases)**	**1EJN+1TNG (36 cases)**
RMSD	0.66 (0.74)	0.64 (0.64)	0.58 (0.63)	0.56 (0.61)	0.87 (0.91)	0.78 (0.78)
GDT_HA	0.67 (0.74)	0.69 (0.68)	0.63 (0.68)	0.71 (0.78)	0.85 (0.92)	0.79 (0.80)
RMSD-s	0.75 (0.83)	0.72 (0.74)	0.76 (0.81)	0.88 (0.92)	0.77 (0.81)	0.65 (0.71)
dRMSD-s	0.68 (0.80)	0.70 (0.78)	0.67 (0.77)	0.87 (0.93)	0.80 (0.83)	0.72 (0.83)
ACS	0.70 (0.75)	0.73 (0.72)	0.69 (0.72)	0.90 (0.88)	0.76 (0.87)	0.82 (0.80)
Seq_Id	0.62 (0.66)	0.56 (0.56)	0.51 (0.58)	0.64 (0.63)	0.87 (0.86)	0.71 (0.76)
RMSD(t)	0.53 (0.53)	0.49 (0.50)	0.58 (0.61)	0.67 (0.65)	0.75 (0.82)	0.38 (0.38)
LGA_RMSD(t)	0.54 (0.55)	0.49 (0.49)	0.58 (0.63)	0.73 (0.70)	0.84 (0.89)	0.43 (0.52)
LGA_S(t)	0.19 (0.23)	0.18 (0.22)	0.21 (0.30)	0.59 (0.51)	0.41 (0.64)	0.34 (0.42)
Seq_Id-s	0.63 (0.66)	0.56 (0.54)	0.56 (0.62)	0.58 (0.61)	0.93 (0.80)	0.76 (0.79)
RMSD-s(t)	0.66 (0.65)	0.67 (0.63)	0.66 (0.68)	0.87 (0.86)	0.85 (0.76)	0.67 (0.55)
MolProbity	0.55 (0.62)	0.50 (0.49)	0.41 (0.47)	0.51 (0.54)	0.86 (0.90)	0.83 (0.77)
Procheck	0.04 (0.08)	0.09 (0.15)	0.06 (0.06)	0.35 (0.41)	0.46 (0.45)	0.59 (0.50)
ProsaII	0.40 (0.48)	0.43 (0.50)	0.34 (0.40)	0.58 (0.67)	0.79 (0.90)	0.35 (0.45)
Verify3D	0.57 (0.63)	0.60 (0.62)	0.47 (0.47)	0.65 (0.72)	0.67 (0.76)	0.63 (0.72)
LGscore	0.49 (0.50)	0.47 (0.47)	0.38 (0.41)	0.65 (0.74)	0.65 (0.82)	0.45 (0.48)
MaxSub	0.45 (0.49)	0.47 (0.45)	0.37 (0.43)	0.59 (0.57)	0.62 (0.82)	0.62 (0.57)
MQscore	0.38 (0.45)	0.44 (0.52)	0.28 (0.38)	0.63 (0.69)	0.60 (0.45)	0.51 (0.54)

Similar trends were obtained for the *mindRMSD* (Table 4.5) and in this case, where the errors associated to the incorrect ranking of the best pose were eliminated, the resulting correlation coefficients are in general higher both for the test set and the subsets. For the 'calculated' indices considered, all the resulting |ρ| values for the whole set are greater than 0.7, and |ρ| = 0.83 and 0.80 were found for the *RMSD-s* and the *dRMSD-s*, respectively. This indicates a high correlation between docking accuracy and both the global adherence of the modelled binding site to the experimental structure and the accuracy of the side-chain conformations in the site.

An investigation of the possibility of developing multivariate models by regression analyses of *dRMSD* and *mindRMSD*, versus all the 'calculated' model quality indices was also performed. From this regression analysis, R^2 coefficients of 0.63 and 0.73 were obtained. Interestingly, the three site quality indices were the most statistically significant in the models obtained (with P-values lower than 0.001), thus confirming that indices of conformity to the native binding site are the most informative in the evaluation of docking results accuracy.

The relationships between docking *dRMSD* (and *mindRMSD*) and each 'predicted' quality index were also investigated. The results of the pairwise correlation analysis for these indices are reported in Table 4.5, both for the test set and for the subsets. The best correlations for the whole test set are those obtained for the global sequence identity with the template, *Seq_Id,* as well as for the indices of conformity of the modelled binding site with the template site, *Seq_Id-s* and *RMSD-s(t)*. The resulting ρ absolute values are higher than 0.6 in both the relationships with docking *dRMSD* and *mindRMSD*, with slightly higher values for the latter index.

The same trend is found when considering the two subsets of T-Coffee and TM-align models: the index that correlates at best with both the *dRMSD* and the *mindRMSD* is the *RMSD-s(t)*, followed by the *Seq_Id* and *Seq_Id-s* for both the subsets and by the *RMSD(t)* and *LGA_RMSD(t)* for the TM-align models.

The ρ values found for these indices on the whole set of models indicate the presence of trends but preclude the possibility of building predictive models of general use. In particular, the plot of *dRMSD vs. Seq_Id* (Fig. 4.12) indicates that the commonly accepted 'rule' that only models with over 50% sequence identity with the template are suitable for docking studies is not reliable. In fact, in many cases for which *Seq_Id* > 50%, docking results with *dRMSD* values from 2 to 8 Å were obtained. Conversely, in some cases acceptable results (*dRMSD* < 2 Å) were obtained by using models with *Seq_Id* < 50%.

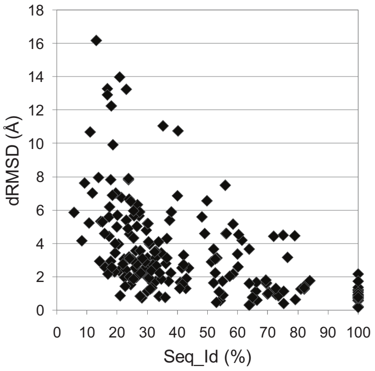


Figure 4.12 – Plot of docking *dRMSD* vs. model-template sequence identity (*Seq_Id*).

Noticeably, for the three quality indices that are better correlated in the whole set with the docking *dRMSD* (and *mindRMSD*) the correlation coefficients are higher in the three homogeneous subsets ($0.6 < |\rho| < 0.9$) than in the whole test set (Table 4.5), as are the correlations with the 'calculated' quality indices. Moreover, in some of the three subsets the correlations with an index of global structural conformity to the template, *RMSD(t),* and the statistical Z scores *MolProbity* and *Verify3D* emerge with comparable ρ values. It can be concluded that in the three subsets all the six most effective indices for model quality prediction (*Seq_Id*, *RMSD(t)*, *Seq_Id-s*, *RMSD-s(t)*, *MolProbity* and *Verify3D*) correlate well with docking accuracy.

It has been observed that the 'predicted' model quality indices that showed low correlation with the 'calculated' indices (Table 4.2), do not correlate with the indices of docking accuracy (Table 4.5). Therefore the ability to predict the quality of a structural model appears to be a necessary pre-requisite, even if not sufficient for predicting the accuracy of the docking results on that model.

Finally, as for the 'calculated' indices, a regression analysis of *dRMSD* and *mindRMSD vs.* all the 'predicted' quality indices was performed to investigate the possibility of developing multivariate models for the relationship between docking and homology modelling accuracy. The *RMSD-s(t)* and *Verify3D* indices were the most statistically significant in the global models (with P-values less than 0.01). The R^2 coefficients resulting from the bivariate models of *dRMSD* and *mindRMSD vs.* these two indices are indeed higher than what obtained for the monovariate models (R^2 = 0.49 and R^2 = 0.48, respectively). However, these models are still only suitable in regression, since they explain about 50% of the variance of the data.

4.1.3 Discussion

Homology models have been increasingly used in ligand-protein docking [51, 56, 59], significantly extending the list of targets available for drug design. Consequently, there is nowadays a great interest in assessing the effects of inaccuracies in protein models on the prediction of protein-ligand interactions [52-54, 56, 140] and in finding general criteria to estimate in advance the accuracy of docking results given the quality of the model [57].

In this work the model features that mostly affect docking accuracy were identified through a large-scale theoretical experiment on a diverse set of ligand-protein complexes.

By studying the relationships between docking accuracy and ‘calculated’ quality indices, as expected, good correlations were found in groups of models of the same protein-ligand complex. For example, for the 16 docking runs performed on different models of 1UKZ (Figure 4.9) the docking dRMSD correlates with all the ‘calculated’ quality indices with $|\rho| > 0.85$. An unexpected result was that good correlations emerged also within groups of different complexes whose proteins are structurally similar (see Table 4.5). As an example, some docking poses for the 1EJN + 1TNG subset are reported in Figure 4.13. The structural superimposition of the two experimental complexes (Fig. 4.13a) confirms a high similarity at the fold level and structurally very well conserved binding sites. The increase of dRMSD of the binding poses correlates well with the decrease of model quality reported in Figures 4.13b-e. Considering the whole subset (36 docking cases), the dRMSD correlates well with all the ‘calculated’ quality indices (see Table 4.5). This agreement was observed in all the three selected subsets of complexes. To this extent, it is remarkable that this result is independent of the similarity of the ligands: in the first case the proteins bind the same ligand (adenosine monophosphate for 2AK3 and 1UKZ); in the second the ligands have similar structures but different functional groups (the all-trans axerophthene in 1FEN differs from the retinoic acid in 1CBS for the presence of a methyl group instead of a carboxyl group); while in the 1EJN + 1TNG set, the ligands have different structures (see caption in Figure 4.13 for details).

When the correlation analysis was extended to the whole test set, where proteins span a large spectrum of structural characteristics (see Table 4.1), lower correlation coefficients were found for indices of global model quality (Table 4.5). This suggests that errors in modelling different folds affect the correlation with docking accuracy in different ways. On the contrary, the correlations of dRMSD with indices of binding site quality are similarly high when considering either the complete set or the single subsets. Also a multivariate regression analysis confirmed that the conformity to the native binding site is the most relevant feature to provide accurate docking results.

In particular, the accuracy in modelling the conformations of the active site sidechains plays an important role in docking into homology models, as shown by the dRMSD-s index. To this extent the ability to include dynamic changes occurring in protein binding sites upon ligand binding is becoming a central issue in molecular docking and many efforts have been made over the past years in developing new docking methods that allow fitting and scoring of flexible ligands in flexible binding sites.

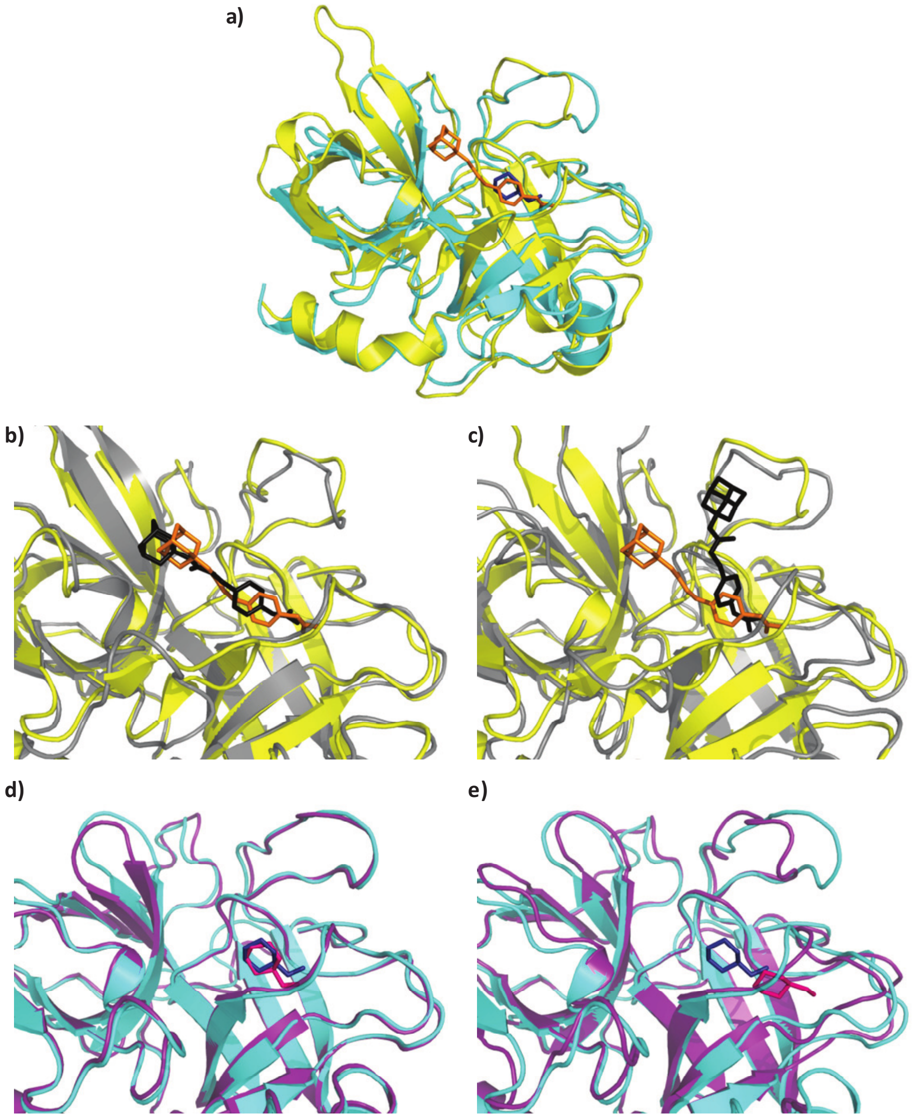


Figure 4.13 – Some docking results for the 1EJN + 1TNG subset of protein-ligand complexes: Urokinase-type plasminogen activator – N-(1-adamantyl)-N'-(4-guanidinobenzyl) urea complex + Trypsin – Aminomethylcyclohexane complex. a) The two experimental structures of the complexes upon structural superimposition of the proteins: 1EJN, protein coloured in yellow, ligand in orange; 1TNG, protein in cyan, ligand in blue; b), c) 1EJN: two binding geometries obtained by docking calculations for protein models of different qualities (models in grey, ligands in black), compared to the experimental geometry (yellow and orange); d), e) 1TNG: two binding geometries obtained by docking calculations for protein models of different qualities (models are coloured in purple and ligands in pink), compared to the experimental geometry (cyan and blue).

These approaches include simulation of the sidechain flexibility during docking, constrained geometrical simulations, ensemble docking to structure sets (from Molecular Dynamics or Monte Carlo methods, rotamer libraries, experimental determinations), or optimization of docked solutions obtained from rigid-receptor docking [30, 147-149]. The results of this work confirm the importance of the development of such strategies also for extending the use of docking methods to homology models.

A wide choice of docking methods is currently available. In order to make this research more useful for the scientific community, the most widely used and tested method for ligand-protein docking [36], AutoDock [29, 106] was selected. Due to limitations imposed by the program when a large conformational space has to be searched [29], it was decided not to employ a recently added option that allows inclusion of protein side-chain flexibility. While this is a more simplified approach to the docking problem, it allows a direct interpretation of the errors associated with the inaccuracies in modelling the active site.

Additionally a difficulty in evaluating the relationships between docking results and model accuracy can arise from the known difficulties of many available docking methods in ranking in a correct way the calculated poses [34, 36]. In this work it was verified that indeed, in some of the cases of the test set, the top ranked poses identified by Autodock reproduced poorly the experimental geometry (Figure 4.10a). When the poses that absolutely better reproduced the experimental binding geometry (mindRMSD) were evaluated (Table 4.5), as expected, better correlations were found. On the other hand, it was observed that the ranking errors were always associated to very limited differences in score (Figure 4.10b). Accordingly, similar conclusions regarding the relationships with the model quality indices were derived for the best geometrical poses as for the best scored poses: there is a higher correlation in homogeneous subsets for global quality indices and a better overall performance of site quality indices.

The analysis of the relationships between docking accuracy and model quality indices derived without any knowledge of the protein experimental structure was performed to verify if an *a priori* prediction of the quality of docking experiments with homology models would be possible. The analysis was performed by employing a set of widely used measures, selected from a large number that have been proposed [89, 141]. The goal of this work was not to provide a comprehensive review of these indices, but to make an assessment of their use in predicting the quality of docking results.

Among the ‘predicted’ indices, the ones showing the best correlations with the ‘calculated’ quality indices (Table 4.2) were considered the most promising for the prediction of docking results accuracy. These include indices that evaluate both the global (Seq_Id, RMSD(t)) and local (Seq_Id-s, RMSD-s(t)) conformity of the model to the template as well as scores (MolProbity [90] and Verify3D [93]) derived by comparison with average properties of known protein structures. However, even if they are effective in predicting the overall quality of the structures, none of these indices comparably correlates with the calculated quality of the binding site. In particular, the target-template sequence identity, that is a widely employed predictive index, gave interesting results. For the test set considered, in many cases the accuracy of the modelled binding site is not directly related to high sequence identity values (Figure 4.6b) and, consistently, a poor predictability of the docking accuracy was found when the correlations between docking dRMSD and Seq_Id were analyzed (see Figure 4.12 and Table 4.5).

Correlation analysis of the dRMSD for the best scored docking pose with all the selected ‘predicted’ indices confirmed these findings (Table 4.5). Acceptable correlations were found only between the dRMSD and the indices of conformity to the template that are mostly related to the calculated model quality, in particular the two indices of binding site quality (Seq_Id-s, RMSD-s(t)). However, for all the ‘predicted’ indices, the mono- and the multivariate regression analyses report weaker correlations than for the ‘calculated’ indices and, due to a lower percentage of explained variance, these models could not be reliable in prediction.

Interestingly, coherently to what was observed for ‘calculated’ indices, in the three homogeneous subsets of proteins the correlations of some ‘predicted’ indices with dRMDS (and mindRMSD) were stronger and all the most effective indices for model quality prediction (Seq_Id, RMSD(t), Seq_Id-s, RMSD-s(t), MolProbity and Verify3D) appeared more suitable for docking accuracy prediction (Table 4.5).

On the basis of this observation, these indices were proposed for predicting docking results accuracy for structurally related proteins.

4.2 A strategy to predict the accuracy of docking poses obtained by using homology models

A relevant outcome of our analysis on the correlations between docking accuracy and model quality indices (see Paragraph 4.1) [66] was the observation that no clear trend exists when evaluating these relationships on heterogeneous groups of proteins. When, instead, proteins belonging to the same CATH superfamily are analyzed, some quality indices showed a certain degree of correlation with docking accuracy. This allowed us to define a strategy for predicting docking results accuracy on homology models.

4.2.1 Computational approach

A workflow of the proposed approach is reported in Figure 4.14.

This consists in: modelling (by homology modelling and ligand-protein docking) a series of complexes whose structure is available in the PDB and whose proteins are homologous to the one that has to be studied; evaluating the docking results by comparison to the native structure of the complex; evaluating the quality of the models. The information produced in this way is used to obtain a multivariate linear regression model that is then employed to predict the accuracy of docking results for the model of interest.

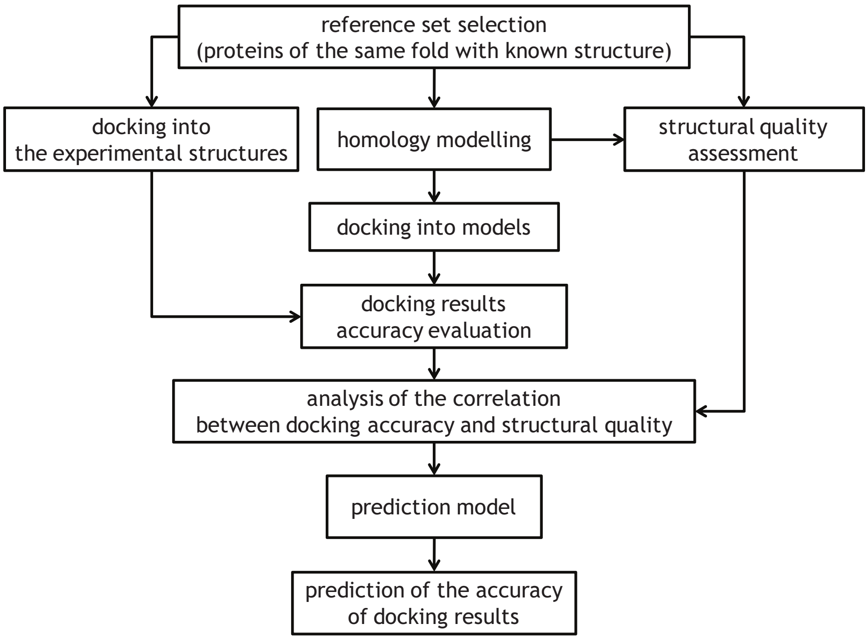

reference set selection
(proteins of the same fold with known structure)
docking into
the experimental structures
homology modelling
structural quality
assessment
docking into models
docking results
accuracy evaluation
analysis of the correlation
between docking accuracy and structural quality
prediction model
prediction of the accuracy
of docking results

Figure 4.14 – Workflow of the proposed prediction strategy.

4.2.2 Application to a study case

The results obtained from the analysis described in Paragraph 4.1 on the subset 1CBS+1FEN were used to develop a multivariate linear regression model. The variables included in the model were the 'predicted' indices with highest correlation ($|\rho| > 0.65$) with docking results accuracy for the subset: *Seq_Id*, *Seq_Id-s*, *RMSD(t)*, *RMSD-s(t)*, *MolProbity* and *Verify3D* (see Table 4.5).

The multivariate model obtained for the subset 1CBS+1FEN (14 points in total) is the following:

dRMSD = - 5.39 + 0.16 Seq_Id – 0.09 Seq_Id-s + 5.07 RMSD(t) + 2.86 RMSD(t)-s – 0.16 MolProbity – 0.67 Verify3D

It is suitable for regression ($R^2 = 0.97$) and for prediction ($Q^2 = 0.84$).

The resulting multivariate model mainly depends on: target-template global sequence identity and global RMSD, thus highlighting that an overall good adherence of the template to the target is important for docking; the sequence identity and RMSD between model and template in the binding site, underlining the importance of a structural description of the binding site similar to the native, that can be modelled at best when it is highly conserved in terms of physico-chemical properties.

The model was used to predict the accuracy of docking results in a new case not included in the original test set, belonging to the same CATH homologous superfamily as 1CBS and 1FEN: the complex between human cellular retinol-binding protein II and retinol (PDB ID: 2RCT, sequence identity equal to 37% and 18% with 1CBS and 1FEN, respectively). Two different homology models were generated for this structure, from two different templates, one at low sequence identity (31% - rat intestinal fatty acid binding protein – PDB ID: 1ICN), and the other at high sequence identity (90% - rat cellular retinol-binding protein II – PDB ID: 1OPA). The sequences of the templates were aligned to the target with T-Coffee, to obtain pair-wise sequence alignments and the two models were generated using Modeller as described in Paragraph 4.1.1. The models were evaluated by calculating *Seq_Id*, *Seq_Id-s*, *RMSD(t)*, *RMSD-s(t)*, *MolProbity* and *Verify3D*.

Some selected 'calculated' and 'predicted' quality indices for the models generated for the cellular retinoic binding protein are reported in Table 4.6.

Table 4.6 – Values of some selected 'calculated' and 'predicted' model quality indices for the two homology models generated for the cellular retinoic binding protein (PDB ID: 2RCT).

Modelname	*RMSD*	*RMSD-s*	*Seq_Id*	*Seq_Id-s*	*RMSD(t)*	*RMSD-s(t)*	*MolProbity*	*Verify3D*
2RCT_1ICN	2.07	2.53	31.3	37.5	0.66	0.78	-15.58	-1.28
2RCT_1OPA	0.90	0.74	90.2	93.8	0.15	0.11	-7.38	-1.61

As expected, the model deriving from the template at higher sequence identity – both globally and in the binding site (2RCT_1OPA) – showed lower RMSD values both with the native overall structure and the binding site. However, it has to be said that the model generated from the template at 30% sequence identity (2RCT_1ICN) is not dramatically different from the native structure, having a global RMSD value of about 2 Å.

The docking of the retinol into the two models was performed using AutoDock, following the procedure described in Paragraph 4.1.1.

In Figure 4.15 the results of the modelling and the docking are reported and compared with the native structure of the complex.

Based on the values of the model quality indices derived for the structures, the docking accuracy was predicted, using the multivariate model, obtaining predicted dRMSD values equal to 5.29 Å for the 2RCT_1ICN model and to 4.38 Å for 2RCT_1OPA. Strikingly, for the two models generated starting from templates at very different sequence identity, also in the binding site, not so different docking results accuracy was predicted.

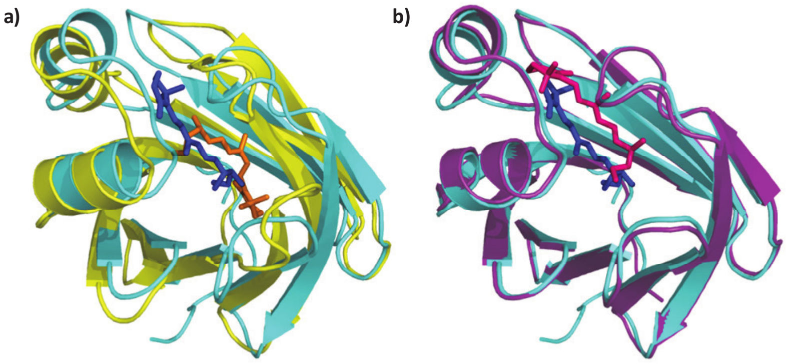


Figure 4.15 – Modelling and docking results for the complex between the human cellular retinol-binding protein II and retinol (PDB ID: 2RCT). a) and b): experimental structure of the complex (in cyan the protein, in blue the ligand), superimposed to the models obtained by using 1ICN (in yellow) and 1OPA (purple) as templates and to the relative docking results (in orange and magenta, respectively).

The results of this prediction, compared to the real docking accuracy, are very interesting: a very small difference (0.37 Å and 0.65 Å for the model built by using 1ICN and 1OPA) is found between the dRMSD calculated by comparison with the known experimental structure and the dRMSD obtained applying the prediction model, as shown in Figure 4.16 and in Table 4.7.

This is a demonstration of the validity of this approach to predict docking results accuracy on the basis of model quality indices.

In summary, these results not only confirmed our observations on the large-scale analysis described in Paragraph 4.1 but also open a way for the prediction of the accuracy of ligand-protein docking results based on model quality indices. This can be of great help in several drug design projects for which no experimental structure of the protein target is available.

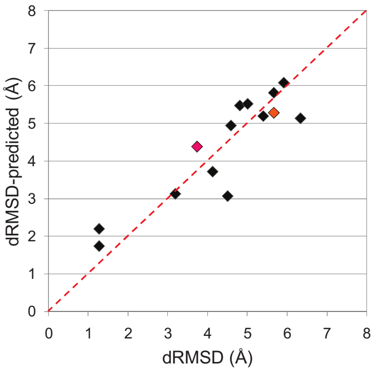


Figure 4.16 – Plot of predicted docking dRMSD (*dRMSD-predicted*) vs. calculated dRMSD (*dRMSD*); in black, values for the models in the subset 1CBS+1FEN, in orange and in magenta, values for the models of 2RCT obtained by using 1ICN and 1OPA as templates, respectively.

Table 4.7 – Values of actual and predicted dRMSD for the docking results on the homology models generated for the cellular retinoic binding protein (PDB ID: 2RCT).

Modelname	Actual dRMSD (Å)	Predicted dRMSD (Å)	Δ (Å)
2RCT_1ICN	5.66	5.29	0.37
2RCT_1OPA	3.73	4.38	0.65

Chapter 5

The use of homology models in protein-protein docking experiments

This chapter illustrates a preliminary analysis performed on a set of modelled and experimental protein structures aimed at assessing the potentialities of the use of homology models in protein-protein docking experiments. Through the selection of templates at different evolutionary distances, a wide range of model quality was spanned. The modelling and the docking steps were set up with the purpose of minimizing errors derived from target-template sequence alignment or docking limitations in sampling the conformational space of protein complexes. The quality of the models obtained was evaluated with a large number of model quality indices, selected also on the basis of the results obtained for an analogous study performed on ligand-protein complexes (see Chapter 4). The preliminary results presented in this chapter indicate that this could be an appropriate and fruitful way to investigate how model quality influences the accuracy of protein-protein docking results.

5.1 Analysis of the relationships between the accuracy of docking results and the quality of protein models

As discussed in the introduction, the recent improvements of homology modelling techniques have greatly extended the possibility of modelling protein structures with increasing accuracy. As discussed in Chapter 4, in the last years this fact has prompted us and other Authors at investigating the possibility of using theoretical protein models for ligand-protein docking and at assessing the strengths and limitations of such applications [51-54, 56, 59, 66, 140]. This subject is assuming increasing relevance also in the protein-protein docking field, as testified by the latest CAPRI rounds, in which is often asked to the participants to model the structure of one or both the interacting molecules [60, 61]. In fact, the possibility of performing protein-protein docking simulations even without the knowledge of the native structure of one or more partners would significantly extend the list of complexes that could be studied, with great repercussions in the fields of biochemistry, medicinal chemistry and similar disciplines.

In particular, the detection of the model features that mainly influence the accuracy of docking results and, consequently, the possibility of predicting *a priori* their degree of confidence (on the basis of standard quality indices or appropriate strategies) would be of great interest. However, until now, no systematic study that investigate the accuracy of protein-protein docking results using homology models has been published.

In this work, a preliminary study was performed to analyze the potentiality of the use of homology models in protein-protein docking. A selection of model quality indices assessing the model-target structural difference (global and at the interface) was used in order to define which are the most relevant model characteristics for obtaining accurate docking results. In addition, indices assessing the model-template differences related to their evolutionary distance, as well as a set of statistical indices of model quality were employed to verify the possibility of predicting docking results accuracy without any knowledge of the native structure.

5.2 Computational approach

5.2.1 Test set of protein-protein complexes and homology modelling strategy

The selected test set is composed by a reference group of X-ray protein structures and a corresponding dataset of theoretical models. As reference group, six recent CAPRI targets were chosen (see Table 5.1 for a brief description of the complexes and Figure 5.1 for a view of the structures). They are representative of different kinds of interactions and include complexes with different kinds of functions [150-153]. They were selected among the CAPRI cases for which the docking method used in this work, HADDOCK [116, 117], in its most recent versions (2.0 and 2.1) [134], showed a very good performance [117, 131], in order to avoid any error deriving from an incapability of the docking method to treat such complexes.

The workflow of this work is reported in Figure 5.2. A traditional homology modelling strategy was employed to generate the modelling dataset: first, candidate templates structures were identified, then target-template alignments were performed and in the end each target was modelled on each of its template structure.

The choice of the programs for each step and the details of the protocol for the traditional homology modelling procedure were based on recent assessments of homology model strategies [21] and template selection strategies [18], in order to reproduce standard homology modelling experiments as well as to span a large range of model quality.

Identification of candidate templates for both the partner proteins of each CAPRI target was performed by sequence similarity search using PSI-BLAST [142] with default parameters until convergence was reached. Each target was searched against a database of all proteins of known structure from the NCBI database.

Among the candidate templates detected in this way, all hits having low statistical significance (BLAST E-value greater than 0.001) or alignment length shorter than 85% of the target sequence were eliminated. Among the remaining hits, templates having from 20% to 80% sequence identity with the corresponding target were selected. In order to obtain an homogeneous sequence identity distribution, if several targets fell in the same 5% identity region, only one was selected (see Table 5.1 for a description of the selected templates).

Table 5.1 – Test cases and selected templates.

CAPRI target	Complex PDB ID (chain ID)	Protein name	Template PDB ID (chain ID)	Seq_Id (%)
T11/T12	1OHZ (A)	Cellulosomal scaffolding protein A	2VN6 (A)	33
			3KCP (A)	62
			1AOH (A)	72
	1OHZ (B)	Endo-1,4-beta-xylanase Y	2VN6 (B)	34
			1DAQ (A)	49
T18	1T6G (C)	Endo-1,4-beta-xylanase I	3B5L (B)	40
			2B42 (B)	44
			1TE1 (B)	45
			1XYN (A)	52
T26	2HQS (C)	Peptidoglycan-associated lipoprotein	2ZVY (A)	23
			2K1S (A)	28
			1R1M (A)	31
			2AIZ (P)	67
T27	2O25 (A)	Ubiquitin-conjugating enzyme E2-25 kDa	3H8K (A)	26
			1YH2 (A)	39
			1TTE (A)	42
	2O25 (C)	SUMO-1-conjugating enzyme UBC9	2GMI (A)	30
			2C4O (B)	35
			1Z3D (A)	40
			2GJD (A)	56
T40	3E8L (A)	Cationic trypsin	1MKW (K)	34
			1FXY (A)	55
			1BZX (E)	65
			1H9H (E)	79
T41	2WPT (A)	Colicin-E2 immunity protein	2K0D (X)	53
			1GXH (A)	60
			3GKL (C)	64
			1FR2 (A)	65
	2WPT (B)	Colicin-E9	2ERH (B)	69
			1MZ8 (B)	70

In order to minimize the errors derived from the alignments between target and template sequences, a structure-structure alignment method, DALILite [83], was employed to generate the alignments for modelling. This was recently assessed as the structural alignment method that shows the best agreement with the NCBI's human-curated Conserved Domain Database (CDD) [154].

Model construction was performed by using MODELLER 9v8 [67], which implements an approach to comparative modelling by satisfying spatial restraints derived from the alignment of the target sequence with the template structure. The method is described in Paragraph 2.1.1. For each alignment, 100 models were generated following the standard MODELLER procedure; the models were ranked by DOPE score [100], as implemented in MODELLER, and the best 10 were retained for docking.

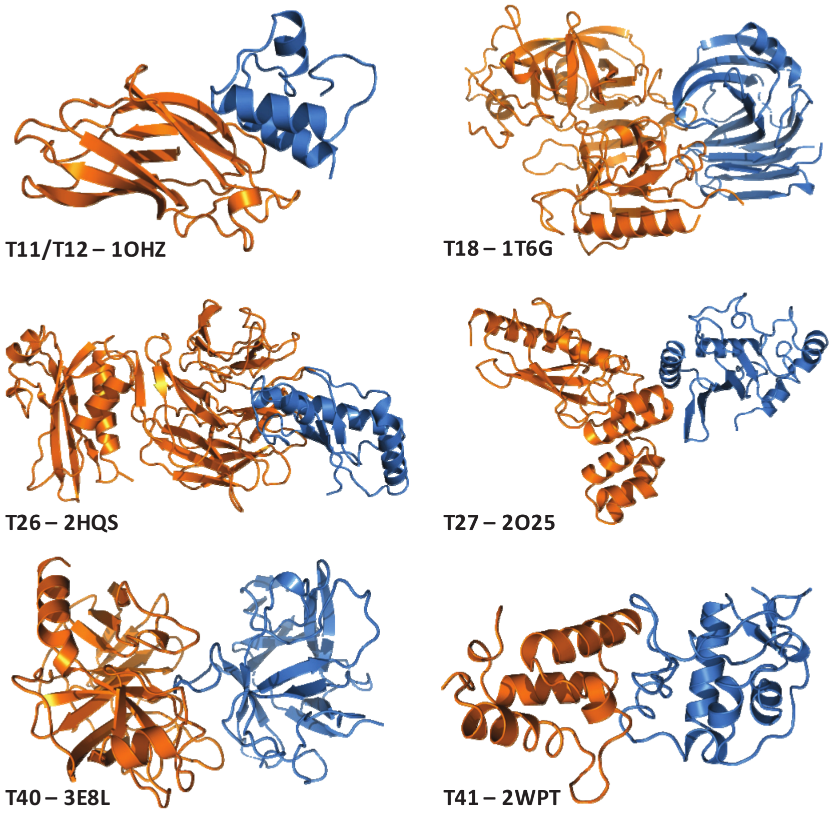


Figure 5.1 – Structures of the complexes in the test set. The proteins are represented in cartoon and different colours discriminate the different molecules in each complex. Under each structure is reported the number of the corresponding CAPRI target and the PDB ID associated to the complex.

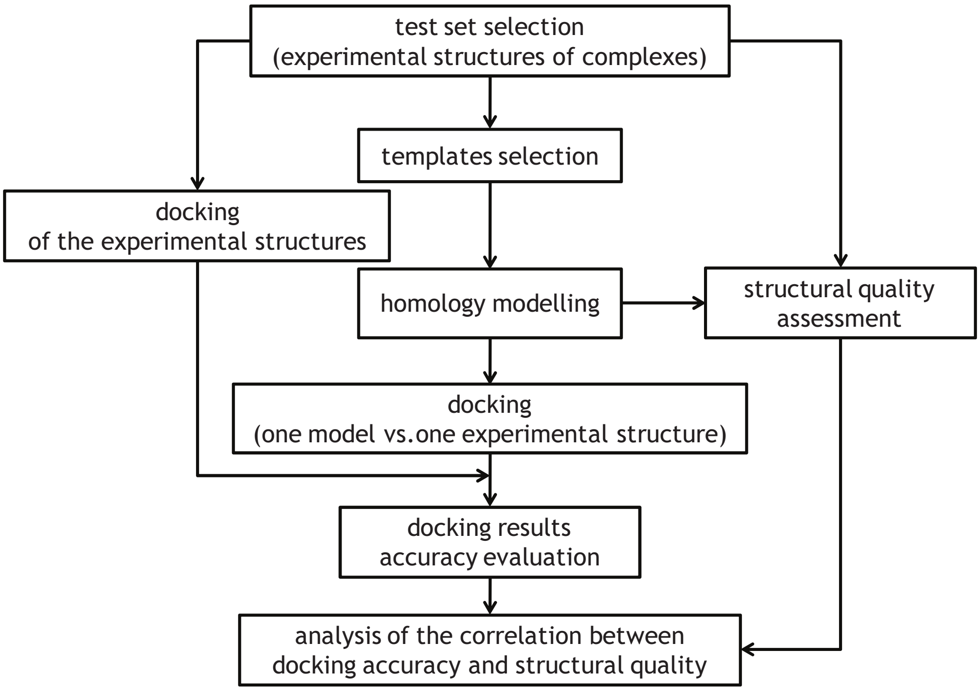

test set selection
(experimental structures of complexes)
templates selection
docking
of the experimental structures
homology modelling
structural quality
assessment
docking
(one model vs.one experimental structure)
docking results
accuracy evaluation
analysis of the correlation between
docking accuracy and structural quality

Figure 5.2 – Workflow of the project.

5.2.2 Model quality indices

As in the ligand-protein docking project described in Chapter 4, models were assessed both by direct comparison to the known native structures ('calculated' indices) and by using indices for model quality estimation and prediction ('predicted' indices). The correlations found in that study [66] drove the selection of the indices to be used in this analysis. In fact, only one index was selected among the quality indices that resulted strongly correlated (for example, the group of indices derived from LGA structural alignments), and the quality indices that were found completely uncorrelated with docking accuracy where discarded beforehand, guessing that they would be uninformative also in this case.

Direct comparison of each model to the corresponding native structure was obtained by structural alignment using the programs DALILite [83], Profit [85] and LGA [84]. These methods are accurately described in Paragraph 2.2, together with all the quality indices mentioned in this paragraph. According to the structural alignment results, the global quality of the models was measured by two indices: the *RMSD* on Cα atoms and *GDT_TS*.

The local quality, which has proven to be crucial for the accuracy of docking results in the ligand-protein docking case, was measured by three indices: *iRMSD_bb*; *iRMSD_sc* and *fnat*. The first is a measure of backbone adherence of the modelled to the native interface; the last two, instead, consider the difference in the interface side-chain geometry between model and target structure.

In the class of 'predicted' indices some indices that evaluate the model-template similarities were considered (see Paragraph. 2.2.2). On the basis of the target-template alignments, the percentage of sequence identity for the whole sequence length (*Seq_Id*) and only for the interface (*Seq_Id-i*) were calculated. The indices *RMSD(t)* and *iRMSD_bb(t)* were also calculated for the model-template structural alignments generated with DALILite.

In addition, a set of 'predicted' indices derived by geometrical analysis or evaluation of the models by statistical potentials was employed (see Paragraph. 2.2.2). These indices are: *MolProbity* [90] and *Verify3D* [93] (from the PSVS server [89]), *DOPE* [100], *TSVMod_RMSD* and *TSVMod_Over* [101] (from the ModEval – Model Evaluation Server by Šali [99]), *Qmean* [103, 104] and *Qmean_Z* (from the Qmean server [102]).

5.2.3 Protein-protein docking calculations

HADDOCK 2.1 [134] was used to perform docking calculations (see also Paragraph 2.4.2). In order to minimize the errors deriving from an insufficient sampling of the conformational space of the complex, the crystallographic interatomic distances ≤ 3.9 Å between the interacting partners were supplied to the docking program as input restraints.

Experimental structures were downloaded from the Protein Data Bank while theoretical models were built as described above. Polar hydrogen atoms were added to each protein structure and the protonation state of histidines was defined using the WHATIF webserver [137]. The docking was performed for each target using both the experimental structures of the two partners, to have a reference of the best results that HADDOCK can achieve with the restraints provided, and the experimental structure of each target with each model of the corresponding partner. As input structures for each docking run, the ensemble of the top 10 models, according to the DOPE score, as implemented in MODELLER 9v8, was given to HADDOCK.

For each docking run, 1100 poses were generated in it0 step and only the best 200 were subjected to refinement in it1 and water steps.

The docking results were evaluated by calculating the i-RMSD with respect to the experimental structure of the complex (see Paragraph 2.4.3). The poses retained for evaluating the relationships between docking results accuracy and model quality were those showing the minimum i-RMSD in each docking run. This choice was made to avoid scoring errors, which could affect the results of the subsequent correlation analyses.

5.3 Results

5.3.1 Model quality variety

For each target and each template 10 models were generated using MODELLER. Out of them, one was selected for model quality assessment, according to docking results. The resulting set includes 30 models.

The quality of each model was at first evaluated by direct comparison with the native structure, employing the 'calculated' indices reported in Paragraph 5.2.2. The modelled set provides a broad spectrum of quality, both global and at the interface, as shown in Figure 5.3.

The distribution of the global *RMSD* values (Fig. 5.3a) shows that the majority of the models have a high adherence to the overall fold of the native structure, with *RMSD* values below 2 Å; some medium-resolution models (*RMSD* values in the range of 2-4 Å) are also present. A minority of the models (7%) show a low similarity with the native structure. This situation reflects also in the distribution of *GDT_TS* values (Fig. 5.3b): the majority of the structures have *GDT_TS* values above 80, thus showing a very good adherence with the native structures; 30% of the models have *GDT_TS* values in the range of 50-80 and only 13% show values below 50 (however, not lower than 40).

The same scenario appears when considering the structural similarity at the interface (Fig. 5.3c-e). The distribution of the *iRMSD_bb* values (Fig. 5.3c) shows that the majority of the models have *iRMSD_bb* values below 2 Å, a minority are in the range of 2-4 Å and only the 13% show *iRMSD_bb* values above 4 Å. When considering the side-chains at the interface (Fig. 5.3d-e), the same overall distribution is found, with the majority of models showing *iRMSD_sc* values below 4 Å and fnat values above 0.7.

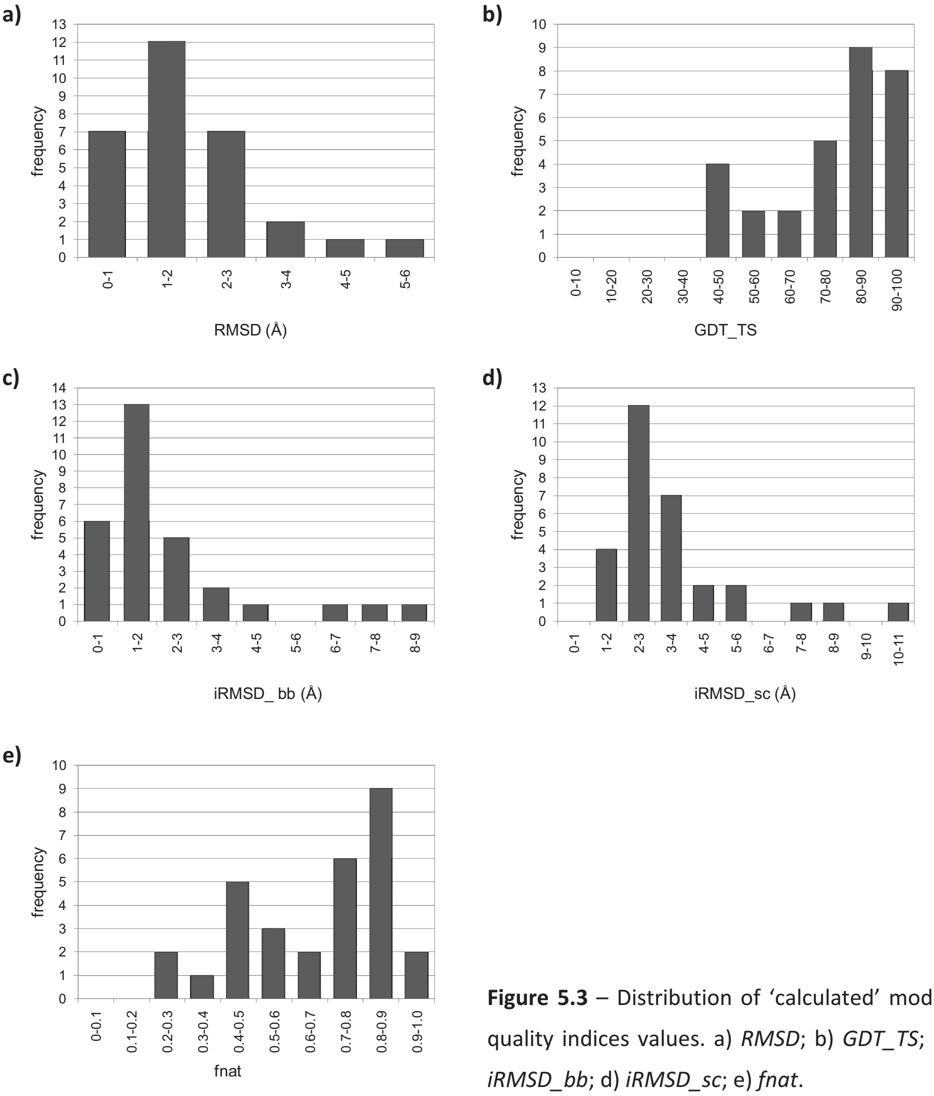


Figure 5.3 – Distribution of 'calculated' model quality indices values. a) *RMSD*; b) *GDT_TS*; c) *iRMSD_bb*; d) *iRMSD_sc*; e) *fnat*.

5.3.2 Correlations between model quality indices

The degree of correlation between the model quality indices was measured by Spearman's rank coefficient, ρ, and the absolute values are reported in Table 5.2.

The high ρ values (all above 0.75) demonstrate that very strong correlations exist within 'calculated' indices. Moreover, no strong difference was found between the correlations within the global quality indices (*RMSD and GDT_TS*) and the interface quality indices (*iRMSD_bb, iRMSD_sc* and *fnat*). Also the degree of correlation between these two sets of indices is similar to intra-set correlations (see for example *GDT_TS* vs. *iRMSD_bb*, that show a ρ = 0.94). This is in disagreement with what found and reported in Chapter 4 for ligand-protein cases and may derive from the fact that protein-protein interface residues are more spread over the protein surface than the binding sites of small molecules, thus the evaluation of interface adherence corresponds also to a description of the fold coherence between model and target structure.

Table 5.2 – Correlations between model quality indices (absolute values of Spearman's rank coefficient). The darker colours represent stronger correlations.

	RMSD	*GDT_TS*	*iRMSD_bb*	*iRMSD_sc*	*fnat*	*Seq_Id*	*Seq_Id-i*	*RMSD(t)*	*iRMSD_bb(t)*	*MolProbity*	*Verify3D*	*DOPE*	*TVSMod_RMSD*	*TVSMod_Over*	*Qmean*	*Qmean_Z*
RMSD	1.00															
GDT_TS	0.87	1.00														
iRMSD_bb	0.95	0.81	1.00													
iRMSD_sc	0.95	0.79	0.94	1.00												
fnat	0.93	0.75	0.94	0.92	1.00											
Seq_Id	0.91	0.85	0.85	0.86	0.84	1.00										
Seq_Id-i	0.79	0.72	0.76	0.73	0.80	0.81	1.00									
RMSD(t)	0.69	0.54	0.58	0.54	0.44	0.71	0.27	1.00								
iRMSD_bb(t)	0.53	0.45	0.50	0.30	0.25	0.50	0.33	0.79	1.00							
MolProbity	0.86	0.76	0.80	0.77	0.85	0.87	0.75	0.63	0.47	1.00						
Verify3D	0.59	0.42	0.56	0.59	0.61	0.55	0.51	0.39	0.26	0.36	1.00					
DOPE	0.68	0.66	0.66	0.71	0.64	0.73	0.38	0.51	0.15	0.65	0.44	1.00				
TVSMod_RMSD	0.76	0.64	0.73	0.77	0.72	0.77	0.56	0.67	0.38	0.59	0.80	0.75	1.00			
TVSMod_Over	0.77	0.67	0.73	0.77	0.73	0.76	0.56	0.57	0.28	0.56	0.80	0.70	0.96	1.00		
Qmean	0.33	0.48	0.22	0.28	0.22	0.34	0.22	0.40	0.17	0.28	0.19	0.28	0.34	0.35	1.00	
Qmean_Z	0.33	0.49	0.23	0.28	0.21	0.34	0.22	0.41	0.17	0.27	0.17	0.27	0.33	0.34	0.99	1.00

Strong correlations were also observed among some ‘predicted’ indices. As expected, the percentage of global sequence identity (*Seq_Id*) is highly correlated with the sequence identity evaluated at the interface (*Seq_Id-i*) (ρ = 0.81). The same is found for the model-template global RMSD (*RMSD(t)*) with the model-template RMSD at the interface (*iRMSD(t)*) (ρ = 0.79). Moreover, the three indices calculated with the ModEval server (*DOPE*, *TSVMod_RMSD* and *TSVMod_Over*) showed a strong correlation. The two indices calculated with the Qmean server (*Qmean* and *Qmean_Z*) were also highly correlated.

Moreover, from the values reported in Table 5.2 it is clear that several ‘predicted’ quality indices are also correlated with ‘calculated’ indices: *Seq_Id*, *Seq_Id-i*, *MolProbity*, *DOPE*, *TSVMod_RMSD* and *TSVMod_Over* have ρ > 0.60 with all the ‘calculated’ indices. In particular, the expected relationship between the target-template sequence identity and the target-model *RMSD* [17] is observed (Fig. 5.4a). In accordance with what observed for ligand-protein complexes (see Chapter 4), models with *Seq_Id* values both greater and lower than 50% can give accurately modelled interfaces (Fig. 5.4b). Anyway, from the plot in Figure 5.4b a *Seq_Id* cut-off for safely obtaining models showing a very accurate description of the backbone geometry at the interface (*iRMSD_bb* < 3 Å), can be detected. This corresponds to *Seq_Id* values above 40%.

The analysis of the correlations between 'calculated' and 'predicted' indices highlighted also the accordance between the RMSD calculated between model and native structure (*RMSD*) and the RMSD predicted using TSVMod (*TSVMod_RMSD*) (ρ = 0.76).

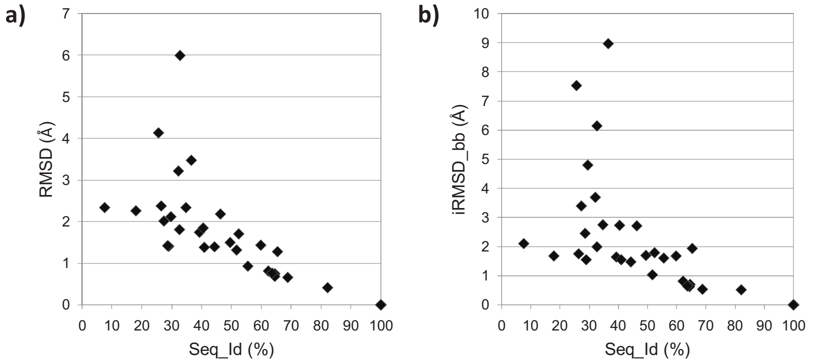


Figure 5.4 – Correlations between *Seq_Id* and: a) *RMSD*; b) *iRMSD_bb*.

5.3.3 Docking results accuracy

The docking experiments were aimed at reproducing the binding geometries corresponding to all the protein-protein complexes in the reference set. To this end, molecular docking calculations were performed both on the protein experimental structures and the associated group of structural models. However, during the docking runs, one of the two interacting partners was always a crystallographic structure. The accuracy of the docking poses obtained was evaluated by calculating the i-RMSD with respect to the native structure of the complexes. The pose showing the lowest i-RMSD for each docking run was taken into account for the subsequent analyses.

The distribution of the *iRMSD* values for the selected docking poses is reported in Figure 5.5. In the majority of the cases, docking results accurately reproduced the experimental binding geometry (*iRMSD* ≤ 2 Å), in 23% of the cases the *iRMSD* was in the range of 2-4 Å, while in a minority of the cases it was above 4 Å. Interestingly, this latter group of results contains only docking solution for one case: the complex between the ubiquitin-conjugating enzyme E2-25 kDa and SUMO-1-conjugating enzyme UBC9 (PDBID: 2O25, corresponding to the CAPRI target no. 27). This had been a controversial case during CAPRI, because of two possible binding modes and small interfaces, that could be due to the crystal packing [60]. Especially the small size of the interface of this complex (870 Å^2) could be the cause of the relatively poor performance of docking for this case, together with a poor adherence to the native interface shown by the models generated for the ubiquitin-conjugating enzyme (Fig. 5.6).

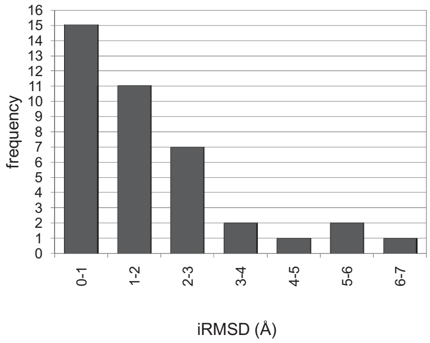


Figure 5.5 – Distribution of *iRMSD* values.

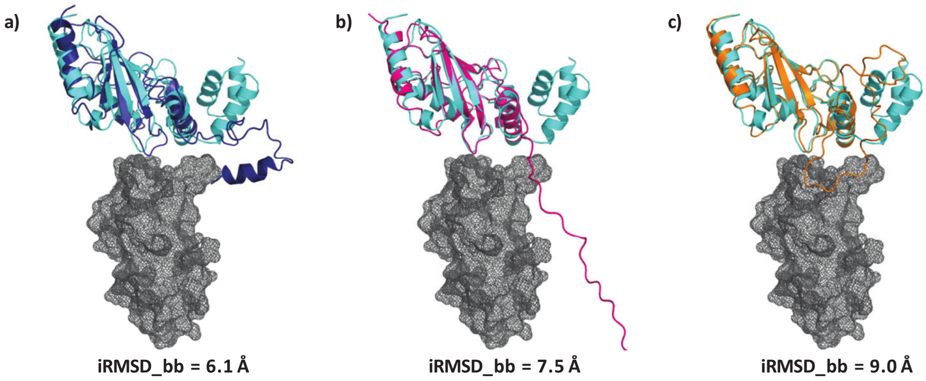


Figure 5.6 – Models generated for the Ubiquitin-conjugating enzyme E2 (PDB ID: 2O25, chain A) (in a) blue, b) magenta and c) orange), superimposed on the native structure (cyan) in complex with the SUMO-1-conjugating enzyme UBC9 (PDB ID: 2O25, chain C) (grey, mesh representation). The corresponding CAPRI target is T27.

The accuracy of docking results was also evaluated by analyzing the percentage of poses with *iRMSD* values ≤ 1 Å ('high-accuracy predictions'), or ≤ 2 Å ('medium-accuracy predictions'), or ≤ 4 Å ('acceptable-accuracy predictions'). The results of this analysis for each test case, are reported in Figure 5.7.

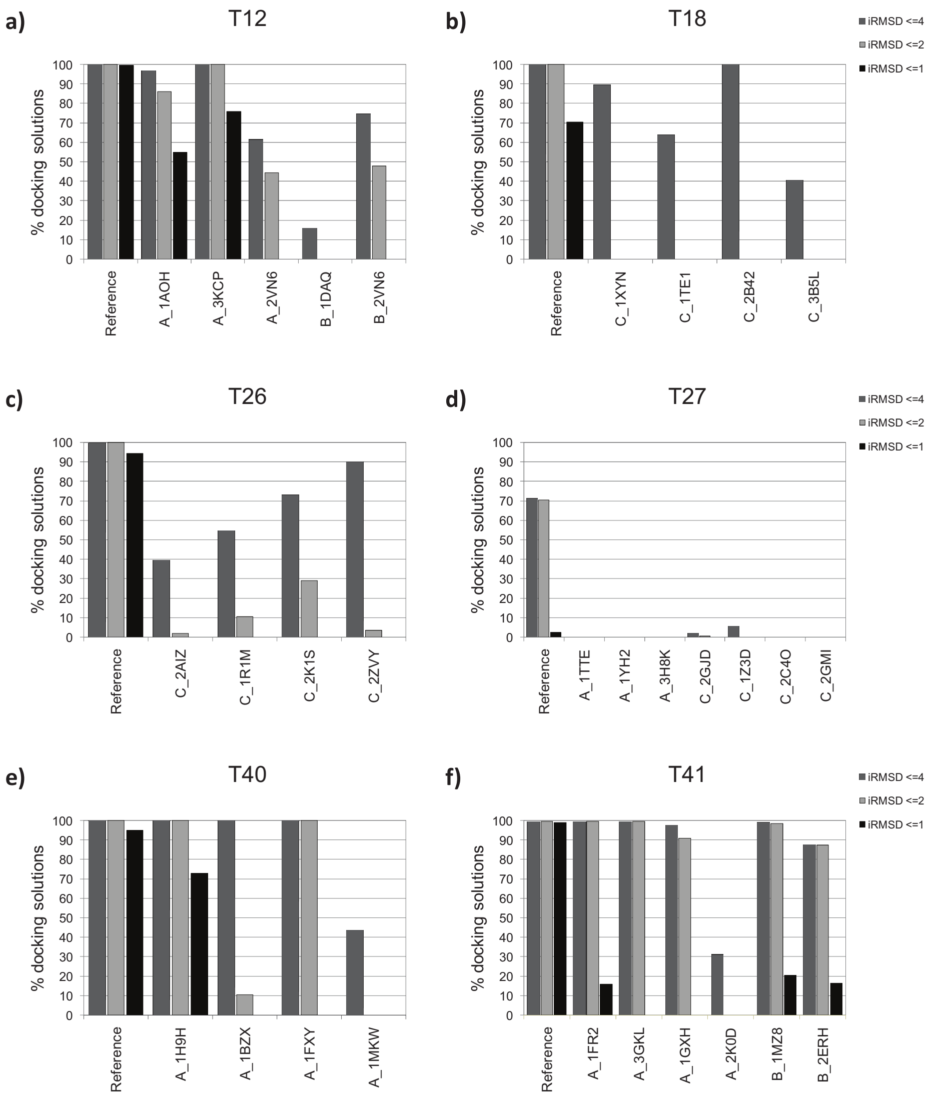


Figure 5.7 – Docking results for each test case: percentage of structures (out of the whole set of docking predictions) that show an iRMSD ≤ 4.0 Å (dark grey), ≤ 2.0 Å (light grey) or ≤ 1.0 Å (black). 'Reference' refers to the docking of the experimental structures; for the docking on the models, the PDB ID of the template used for modeling is reported under the corresponding histogram.

As expected, high-accuracy predictions are more difficult to find than the others, even when both the docked partners are experimental structures. However, for some of the test cases (target 12, target 40 and target 41), a great amount of medium-accuracy solutions and even some high-accuracy predictions are found among the generated docking poses. For target 18 and 26 mostly acceptable solutions are found for the docking runs that involved models. For target 27, in accordance with the high values of *iRMSD* found, almost no acceptable solutions were generated during the docking on homology models.

5.3.4 Relationships between model quality and docking results accuracy

The Spearman's rank coefficients for the correlation analysis of the 'calculated' and 'predicted' indices with docking accuracy are reported in Table 5.3. Some examples of correlation plots between *iRMSD* and 'calculated' and 'predicted' indices are reported in Figure 5.8 and 5.9, respectively.

Strong correlations were found between *iRMSD* and all the 'calculated' indices. Interestingly, the lowest ρ values (|ρ| = 0.81) are those obtained for the indices of model-target adherence of interface side-chain geometry (*iRMSD_sc* and *fnat*). This confirms the ability of HADDOCK in treating the interface side-chain flexibility and suggests that errors in side-chain geometry can be adjusted by HADDOCK flexible refinement.

Table 5.3 – Correlations between iRMSD and model quality indices (absolute values of Spearman's rank coefficient).

'Calculated' indices	iRMSD	'Predicted' indices	iRMSD
RMSD	0.87	*Seq_Id*	0.78
GDT_TS	0.87	*Seq_Id-i*	0.64
iRMSD_bb	0.90	*RMSD(t)*	0.52
iRMSD_sc	0.81	*iRMSD_bb(t)*	0.53
fnat	0.81	*MolProbity*	0.75
		Verify3D	0.46
		DOPE	0.62
		TVSMod_RMSD	0.61
		TVSMod_Over	0.59
		Qmean	0.22
		Qmean_Z	0.23

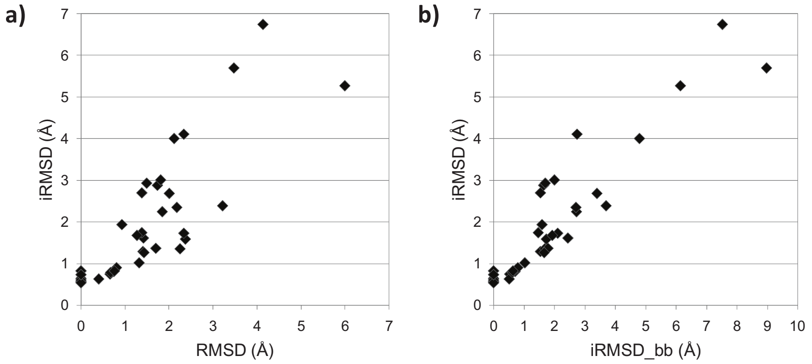


Figure 5.8 – Correlation between *iRMSD* and some examples of 'calculated' indices: a) *RMSD*; b) *iRMSD_bb*.

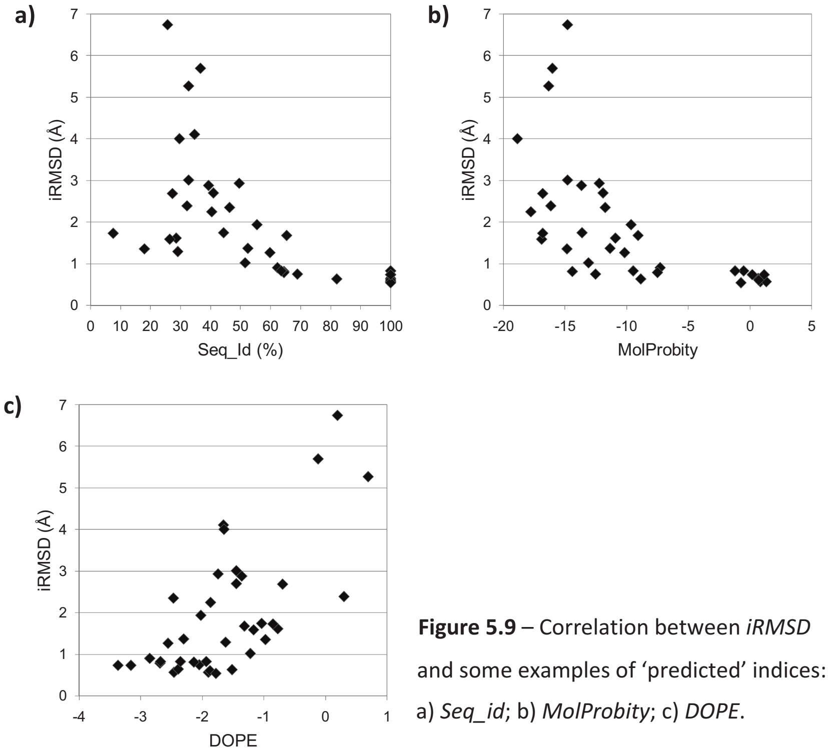


Figure 5.9 – Correlation between *iRMSD* and some examples of 'predicted' indices: a) *Seq_id*; b) *MolProbity*; c) *DOPE*.

Satisfactory correlations between 'predicted' indices and docking results accuracy were also obtained. High ρ values were found for *Seq_Id* (|ρ| = 0.78) and *Seq_Id-i* (|ρ| = 0.64). Strikingly, for the test cases here considered, the relation between *Seq_Id* and *iRMSD* values (Fig. 5.9a) allows to detect a cut-off for safely obtaining very accurate docking results (*iRMSD* < 3 Å). Similarly to what observed in Paragraph 5.3.2 for the correlation between *Seq_Id* and *i_RMSD_bb*, this corresponds to *Seq_Id* > 40%. If this observation will be valid also for a broader and statistically significant test set, the cut-off of 40% sequence identity could become a very useful guide for the selection of templates for modelling structures to be used in protein-protein docking experiments, if a safe docking is strongly required. This does not implies, however, that only non-accurate docking results are obtained from models deriving from templates with *Seq_Id* < 40%, as demonstrated by the plot in Figure 5.9a, where very good results are reported even for *Seq_Id* values < 20%. In addition to this, surprisingly, good correlations were obtained also for indices that do not take into account the model-template similarity: *MolProbity* ($|\rho| = 0.75$), *DOPE* ($|\rho| = 0.62$) and *TSVMod_RMSD* ($|\rho| = 0.61$).

In Figure 5.10 an example of such correlations within a test case is reported. The docking results for the experimental structures of the ligand and receptor for target 40, together with the results of docking calculations of two models at different accuracy generated for the trypsin are shown in the figure. As reported, higher target-template percentage of identity, values of MolProbity clashscore nearer to 0 and lower DOPE scores (which indicate good models) correspond to more accurate docking results.

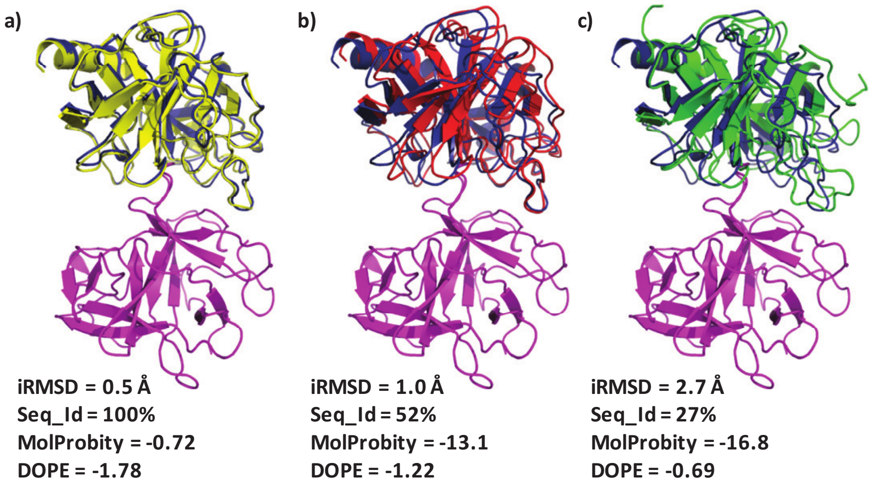


Figure 5.10 – Docking results for the Serine proteinase inhibitor A in complex with Cationic trypsin (PDB ID: 3E8L, CAPRI T40). In blue and in magenta, the experimental structure of the complex. Superimposition of the best docking pose for: a) the trypsin experimental structure (yellow); b) and c) two models generated for the trypsin (red and green, respectively). The corresponding values of *iRMSD* for the docking results and *Seq_Id*, *MolProbity* and *DOPE* for the models are reported.

5.4 Discussion and future perspectives

In the last years, the use of homology models in ligand-protein docking has significantly extended the possibility of studying drug-protein interactions for medicinal chemistry purposes [51, 56, 59]. This has not yet happened for protein-protein docking, for which some limitations at both the sampling and scoring stage still exist. The increasing use of homology models in CAPRI rounds [60, 61], however, indicates the need of extending protein-protein docking calculations also to modelled structures. This would extend the list of protein-protein interactions that could be studied, but will also require some guidelines that state to which extent accurate docking results can be achieved when using theoretical models instead of experimental structures.

In this work, a preliminary analysis was performed to assess which model features are the most crucial in determining the accuracy of docking results. Moreover, a set of model quality indices was considered to assess if it is possible to predict docking results accuracy on the basis of a standard analysis of model quality.

As expected, strong correlations were found between docking accuracy and 'calculated' indices, both of global and local (at the interface) quality. It was interesting to verify that the use of a docking method able to handle the flexibility of the protein interface resulted in a little weaker correlation between *iRMSD* and indices of adherence of interface side-chain geometry (*iRMSD_sc*, *fnat*), compared to the ones found for global RMSD or interface backbone geometry (*iRMSD_bb*). In fact, it is conceivable that the ability of HADDOCK in dealing with flexibility could help in minimizing errors deriving from a poor accuracy in modelling the side-chains at the interface, thus leading to accurate docking results.

Among the 'predicted' indices, the ones showing the best correlations with docking results accuracy belong to different classes of model quality indices. Some are indices of global and local conformity of the template to the target (*Seq_Id* and *Seq_Id-i*); their strong correlations with docking results accuracy confirm the importance of the choice of the good template [18] for modelling. In particular, from the preliminary analysis here performed, it seems that a first guideline to obtain safely accurate docking results on homology models could be to chose templates with sequence identity with the target equal or above 40%. Some other 'predicted' indices that show high correlations with docking results accuracy are scores derived by comparison with average properties of known protein structures (*MolProbity* [90], *TSVMod_RMSD* and *TSVMod_Over* [101]) or

by an energetic analysis of the models (*DOPE* [100]); their high correlation values with *iRMSD* suggest that they may be used for predicting docking results accuracy, even without the knowledge of the native structure of the target protein.
However, the set of models used for this work is too small and therefore not statistically significant to develop a reliable approach for predicting the accuracy of protein-protein docking results.

These preliminary results, obtained for docking calculations driven by crystallographic restraints and for models generated starting from structure-structure alignments in order to minimize docking and modelling errors, will be the basis for future work. The test set will be enlarged, different alignment strategies will be used and different restraints will be exploited to guide the docking. This will lead to a larger and more comprehensive analysis of the problem and could result in a strategy to *a priori* predict docking results accuracy on the basis of model quality indices.

Chapter 6

Conclusions and future perspectives

The principal aim of this thesis was to overcome the limitations of the sampling stage of ligand- and protein-protein docking approaches, exploring the potentialities of combining different computational techniques to broaden the possibility of predicting the structure of protein complexes. In particular, in the protein-protein docking field, I combined two docking methods, each belonging to one of the two mainly used sampling approaches, with the aim of obtaining an algorithm that could overcome the limitations of the two and sum their strengths. Moreover, both for ligand- and protein-protein docking, I focused on the topic of docking calculations on homology models, assessing if it is possible to *a priori* predict the accuracy of docking results on the basis of the evaluation of the model quality.

These aims were motivated by two facts. The first is that both ligand- and protein-protein docking have recently become protagonists in molecular modelling, as they are valid tools for the prediction of the structure of protein complexes. In addition, the recent improvements in homology modelling techniques and the development of model repositories, that nowadays make models available for a large community, have prompted the use of protein models for docking calculations. Therefore, in the last years it raised the need of both developing more and more reliable tools for molecular modelling and defining some guidelines to obtain accurate prediction of the structure of protein complexes. These goals are particularly relevant when only poorly accurate information (e.g. no interface indication for protein-protein docking, or protein models instead of experimental structures) is available.

6.1 The combination of different approaches as a powerful tool to improve the initial search stage of protein-protein docking methods

In Chapter 3 I illustrated the development and performance of ZADDOCK, a combination of ZDOCK and HADDOCK, which exploits and merges the fast rigid-body search performed by ZDOCK and the accurate flexible refinement in explicit solvent of HADDOCK. The analysis of ZADDOCK performance on a wide test set, representing different types of complexes and of interactions, showed that accurate docking results can be obtained by this new method, thus achieving two goals. The first was the possibility of using HADDOCK without the need of any experimental data to guide the sampling step, in order to study complexes for which no experimental information is available and bioinformatics interface prediction fails. The second was to obtain an accurate description of the intermolecular interactions occurring in protein complexes, which is a key information to drive subsequent experimental work.

The quality of ZADDOCK results indicates that the strategy of combining different computational techniques is a very promising avenue for the development of new docking approaches, which will be aimed at coping with the most difficult docking cases. These are the complexes for which big conformational changes occur upon binding or no information about the interface is available.

Since each class of sampling strategy implemented in the different docking methods is optimal for a specific class of docking problems, in fact, it is conceivable that a combination of several protein-protein docking approaches could help to reach improved results. For example, as it was implemented in ZADDOCK, FFT docking methods could be used to generate in a fast way a series of initial docking poses, to be refined with EM algorithms. Another refinement procedure could be, for example, the combination of an EM search that treats the backbone as flexible with MC search for the optimization of the side-chains positions at the interface.

Moreover, in order to cope with cases for which significant conformational changes occur upon binding, computational techniques that predict protein flexibility [155-160] and generate an ensemble of possible conformations for the flexible structures could be combined with an available docking method in an integrated cross-docking approach. The limits of such an approach would certainly be the CPU time requested for calculations and the number of false positives generated, which, currently, scoring functions would not be able to distinguish from the real near-native solutions. Thus, the development of such methods would require also faster search algorithms as well as more accurate scoring functions. Another trend, opened by Wolfson and collaborators with the development of FlexDock [161], is the combination of hinge prediction with the docking of the rigid parts of the flexible molecule followed by the building of consistent configurations of the entire protein from these candidate parts.

On the other hand, as it has been demonstrated in CAPRI experiments, accurate indications about the interface are a very valuable guide for protein-protein docking, having great influence on the accuracy of docking results. Unfortunately, for some cases of interest, no experimental data suggesting which are the residues involved in the interaction are available and bioinformatics interface prediction are not reliable. This could be due, for example, to the variety of interacting partners and of binding modes of these proteins or to a lack of known homologues on which a statistic could be performed. In such cases, the use of fast docking algorithms (for example, FFT approaches) could be of some help for interface prediction. In fact, provided that the right conformation for binding is given to those methods, their ability in finding shape complementary regions allows them to determine with a certain accuracy which parts of the surface of both the interacting partners are the most probable to be at the interface. An example of such strategy was published very recently by Weng and collaborators: they exploited the fast search of their docking method, ZDOCK, for interface prediction in the latest CAPRI rounds, obtaining reliable predictions for the majority of the complexes analyzed [162]. In addition, the observation made in the latest CAPRI assessment that some docking methods show a better performance in interface prediction than the average of the standard bioinformatics interface prediction approaches [133], encourages the future development of such strategies.

6.2 The use of homology models for docking calculations: hints to obtain accurate results

In Chapter 4 and Chapter 5 it was investigated how the quality of homology models, used instead of experimental structures in ligand- and protein-protein docking calculations, influences the accuracy of docking results. The final aim of these analyses was to assess the possibility of *a priori* predicting the accuracy of docking results on the basis of model quality. This was demonstrated with the development of a prediction strategy for ligand-protein docking, and will be the next step for protein-protein docking, after a broadening of the test set to allow a statistically significant analysis of the data.

The trends found in the two docking fields are quite different. In fact, while in ligand-protein docking evident correlations between docking results accuracy and model quality were found especially in structurally similar cases, the preliminary results found for the protein-protein docking simulations indicated stronger and more general correlations. This allows to speculate that a general rule could be found for predicting *a priori* the accuracy of docking results, on the basis of some specific model quality indices. The difficulty in finding general relationships between docking results and model quality for ligand-protein docking experiments reflects what was already observed by other Authors

in terms of relations between enrichment and target-template sequence identity [52-58] and explains why until now no general rule for predicting docking results accuracy on the basis of model quality is available.
However, on the basis of the results obtained in this thesis, it is conceivable that a promising avenue to obtain strong general correlations, and consequently a general prediction rule, in ligand-protein docking, could be the development of knowledge-based potentials for model quality assessment derived from specific information of homologous proteins, as for example the one proposed by Panjkovich et al. [163].

Another difference between the two cases is the crucial importance, found in ligand-protein docking results, of the geometrical reproduction of the native binding region, which did not show the same influence on the accuracy of the results in protein-protein docking. This observation clearly reflects the difference in treating the flexibility of the binding region between the two docking approaches. While, in fact, most protein-protein docking algorithms can treat the proteins as flexible with accurate results, this is still a far achievement for ligand-protein docking methods [29, 30, 147, 164], whose results, consequently, strongly depend on an accurate description of the binding site geometry.
The possibility of overcoming this limitation would be of great interest for the molecular modelling community. Therefore, recently most ligand-protein methods have incorporated algorithms able to treat the side-chains in the receptor binding site as flexible. Examples of these approaches are: Monte Carlo search, rotamer libraries, ensemble docking into sets of structures derived from MD trajectories or from NMR structure determinations. Other methods do not explicitly include flexibility, but make use of soft potentials. However, each different approach is affected by limitations in computational performance or in accuracy. Therefore, a new tendency was proposed also in this field, which is to combine several strategies in order to improve the reliability of docking into flexible proteins [30].

Another way of improving docking results accuracy for homology models could be to generate different homology models starting from different templates and use them in cross-docking calculations. This would minimize the errors deriving from template selection and increase the possibility of a good description on the binding region, thanks to the variability of the structures submitted to docking, as suggested also by Fan et al.

for ligand-docking [57]. Also in this case, good results could be obtained from such a strategy only if scoring functions would be implemented at the point that they can discriminate near-native docking solutions from the others. Otherwise, this approach could be used in combination with a strategy to *a priori* predict the accuracy of docking results on homology models, which could indicate the most suitable models to obtain accurate docking results.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000; 28: 235-242.
2. Pioletti M, Schlünzen F, Harms J, Zarivach R, Glühmann M, Avila H, Bashan A, Bartels H, Auerbach T, Jacobi C, Hartsch T, Yonath A, Franceschi F. Crystal structures of complexes of the small ribosomal subunit with tetracycline, edeine and IF3. EMBO J 2001; 20: 1829-1839.
3. Clore GM, Gronenborn AM. NMR structure determination of proteins and protein complexes larger than 20 kDa. Curr Opin Chem Biol 1998; 2: 564-570.
4. Joachimiak A. High-throughput crystallography for structural genomics. Curr Opin Struct Biol 2009; 19: 573-584.
5. Terwilliger TC, Stuart D, Yokoyama S. Lessons from structural genomics. Annu Rev Biophys 2009; 38: 371-383.
6. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003; 31: 365-370.
7. The UniProt Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res 2007; 35: D193-D197.
8. Stein L. Genome annotation: from sequence to biology. Nat Rev Genet 2001; 2: 493-503.
9. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181: 223-230.
10. Chothia C. Proteins. One thousand families for the molecular biologist. Nature 1992; 357: 543-544.
11. Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families. Protein Eng 1993; 6: 485-500.
12. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991; 253: 164-170.
13. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature. 1992 Jul 2;358(6381):86-89.
14. Bystroff C, Simons KT, Han KF, Baker D. Local sequence-structure correlations in proteins. Curr Opin Biotechnol 1996;7: 417-421.
15. Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci U S A 1996; 93:5814-5818.
16. Krieger E, Nabuurs SB, Vriend G. Homology modeling. In: Bourne PE, Weissig H. Structural Bioinformatics, Wiley-Liss, New Jersey, 2003.
17. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986; 5: 823-826.
18. Sadowski MI, Jones DT. Benchmarking template selection and model quality assessment for high-resolution comparative modeling. Proteins 2007; 69: 476-485.

19. Larsson P, Wallner B, Lindahl E, Elofsson A. Using multiple templates to improve quality of homology models in automated homology modeling. Protein Sci 2008; 17: 990-1002.
20. Cozzetto D, Tramontano A. Relationship between multiple sequence alignments and quality of protein comparative models. Proteins 2005; 58: 151-157.
21. Dalton JA, Jackson RM. An evaluation of automated homology modelling methods at low target template sequence similarity. Bioinformatics 2007; 23: 1901-1908.
22. Kihara D, Chen H, Yang YD. Quality assessment of protein structure models. Curr Protein Pept Sci 2009; 10: 216-228.
23. Kryshtafovych A, Fidelis K. Protein structure prediction and model quality assessment. Drug Discov Today 2009; 14: 386-393.
24. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins 1995; 23: ii-v.
25. Cozzetto D, Kryshtafovych A, Tramontano A. Evaluation of CASP8 model quality predictions. Proteins 2009; 77: 157-166.
26. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002; 47: 409-443.
27. Leach AR, Shoichet BK, Peishoff CE. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. J Med Chem 2006; 49: 5851-5855.
28. Dias R, de Azevedo WF Jr. Molecular docking algorithms. Curr Drug Targets 2008; 9: 1040-1047.
29. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 2009; 30: 2785-2791.
30. B-Rao C, Subramanian J, Sharma SD. Managing protein flexibility in docking and its applications. Drug Discov Today 2009; 14: 394-400.
31. Kontoyianni M, McClellan LM, Sokol GS. Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 2004; 47: 558-565.
32. Perola E, Walters WP, Charifson PS. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 2004; 56: 235-249.
33. Cummings MD, Des Jarlais RL, Gibbs AC, Mohan V, Jaeger EP. Comparison of automated docking programs as virtual screening tools. J Med Chem 2005; 48: 962-976.
34. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem 2006; 49: 5912-5931.
35. Plewczynski D, Lazniewski M, Augustyniak R, Ginalski K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J Comput Chem 2010; DOI: 10.1002/jcc.21643.
36. Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: current status and future challenges. Proteins 2006; 65: 15-26.
37. Tuccinardi T. Docking-based virtual screening: recent developments. Comb Chem High Throughput Screen 2009; 12: 303-314.

38. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. Critical Assessment of PRedicted Interactions. CAPRI: a Critical Assessment of PRedicted Interactions Proteins. 2003; 52: 2-9.
39. Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. Proteins 2010; 78: 3073-3084.
40. Janin J. Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 2010; 6: 2351-2362.
41. Moreira IS, Fernandes PA, Ramos MJ. Protein-protein docking dealing with the unknown. J Comput Chem 2010; 31: 317-342.
42. van Dijk AD, Boelens R, Bonvin AM. Data-driven docking for the study of biomolecular complexes. FEBS J 2005; 272: 293-312.
43. Melquiond AS, Bonvin AM. Data-driven docking: using external information to spark the biomolecular rendez-vous. In: Protein-protein complexes: analysis, modelling and drug design. Edited by M. Zacharias, Imperial College Press, 2010. p 183-209.
44. Vajda S, Kozakov D. Convergence and combination of methods in protein-protein docking. Curr Opin Struct Biol 2009; 19: 164-170.
45. Kastritis PL, Bonvin AM. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 2010; 9: 2216-2225.
46. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction-Round VII. Proteins 2007; 69: 3-9.
47. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction - Round VIII. Proteins 2009; 77: 1-4.
48. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 2009; 37: D387-D392.
49. Arnold K, Kiefer F, Kopp J, Battey JN, Podvinec M, Westbrook JD, Berman HM, Bordoli L, Schwede T. The Protein Model Portal. J Struct Funct Genomics 2009; 10: 1-8.
50. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, Webb B, Greenblatt D, Huang CC, Ferrin TE, Šali A. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 2004; 32: D217-D222.
51. Cavasotto CN, Phatak SS. Homology modeling in drug discovery: current trends and applications. Drug Discov Today 2009; 14: 676-683.
52. Diller DJ, Li R. Kinases, homology models, and high throughput docking. J Med Chem 2003; 46: 4638-4647.
53. McGovern SL, Shoichet BK. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 2003; 46: 2895-2907.
54. Oshiro C, Bradley EK, Eksterowicz J, Evensen E, Lamb ML, Lanctot JK, Putta S, Stanton R, Grootenhuis PD. Performance of 3D-database molecular docking studies into homology models. J Med Chem 2004; 47: 764-767.
55. Kairys V, Fernandes MX, Gilson MK. Screening drug-like compounds by docking to homology models: a systematic study. J Chem Inf Model 2006; 46: 365-379.

56. Ferrara P, Jacoby E. Evaluation of the utility of homology models in high throughput docking. J Mol Model 2007; 13: 897-905.
57. Fan H, Irwin JJ, Webb BM, Klebe G, Shoichet BK, Šali A. Molecular docking screens using comparative models of proteins. J Chem Inf Model 2009; 49: 2512-2527.
58. Novoa EM, de Pouplana LR, Barril X, Orozco M. Ensemble docking from homology models. J Chem Theory Comput 2010; 6: 2547-2557.
59. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. Drug Discov Today 2004; 9: 659-669.
60. Janin J. The targets of CAPRI Rounds 13-19. Proteins 2010; 78: 3067-3072.
61. Janin J. The targets of CAPRI rounds 6-12. Proteins 2007; 69: 699-703.
62. Lorenzen S, Zhang Y. Monte Carlo refinement of rigid-body protein docking structures with backbone displacement and side-chain optimization. Protein Sci 2007; 16: 2716-2725.
63. Kozakov D, Schueler-Furman O, Vajda S. Discrimination of near-native structures in protein-protein docking by testing the stability of local minima. Proteins 2008; 72: 993-1004.
64. Pierce B, Weng Z. A combination of rescoring and refinement significantly improves protein docking performance. Proteins 2008; 72: 270-279.
65. Liang S, Wang G, Zhou Y. Refining near-native protein-protein docking decoys by local resampling and energy minimization. Proteins 2009; 76: 309-316.
66. Bordogna A, Pandini A, Bonati L. Predicting the accuracy of protein-ligand docking on homology models. J Comput Chem 2011; 32: 81-98.
67. Šali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 1993; 234: 779-815
68. Fiser A, Do RK, Šali A. Modeling of loops in protein structures. Protein Sci 2000; 9: 1753-1773.
69. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000; 29: 291-325.
70. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007; 5: 17.
71. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008; 9: 40-47.
72. MacKerell Jr AD, Bashford D, Bellott M, Dunbrack Jr RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher III WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998; 102: 3586–3616.
73. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 2004; 101: 7594-7599.
74. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983; 22: 2577-2637.
75. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999; 292: 195-202.

76. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48: 443-453.

77. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002; 48: 192-201.

78. Chen H, Zhou HX. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 2005; 33: 3193-3199.

79. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004; 25: 865-871.

80. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005; 33: 2302-2309.

81. Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL 3rd. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 2000; 41: 86-97.

82. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003; 12: 2001-2014.

83. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993; 233: 123-138.

84. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003; 31: 3370-3374.

85. Martin ACR, Porter CT. http://www.bioinf.org.uk/software/profit/

86. McLachlan AD. Rapid Comparison of Protein Structres. Acta Cryst A 1982; 38: 871-873.

87. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007; 69: 38-56.

88. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992; 89: 10915-10919.

89. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. Proteins 2007; 66: 778-795.

90. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 2003; 50: 437-450.

91. Laskowski R A, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 1993; 26: 283-291.

92. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993; 17: 355-362.

93. Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992; 356: 83-85.

94. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003; 12: 1073-1086.

95. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A. A study of quality measures for protein threading models. BMC Bioinformatics 2001; 2:5.

96. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000; 16: 776-785.

97. McGuffin LJ. The ModFOLD server for the quality assessment of protein structural models. Bioinformatics 2008; 24: 586-587
98. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004; 57: 702-710.
99. Eramian D, Shen MY, Melo F, Pieper U, Webb B, Šali A. http://modbase.compbio.ucsf.edu/evaluation/
100. Shen MY, Šali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006; 15: 2507-2524.
101. Eramian D, Eswar N, Shen MY, Šali A. How well can the accuracy of comparative protein structure models be predicted? Protein Sci 2008; 17: 1881-1893.
102. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res 2009; 37: W510-W514.
103. Benkert P, Tosatto SC, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008; 71: 261-277.
104. Benkert P, Schwede T, Tosatto SC. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Struct Biol 2009; 9: 35.
105. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 2010 DOI: 10.1093/bioinformatics/btq662.
106. Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. J Comput Chem 2007; 28: 1145-1152.
107. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 1998; 19: 1639-1662.
108. Solis FJ, Wets RJB. Minimization by random search techniques. Math Oper Res 1981; 6: 19–30.
109. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, AlagonaG, Profeta Jr S, WeinerP. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984; 106: 765-784.
110. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985; 28: 849-857.
111. Nissink JW, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein-ligand interaction. Proteins 2002; 49: 457-471.
112. Clark M, Cramer III RD, Van Opdenbosch N. Validation of the general purpose tripos 5.2 force field. Proteins 1989; 10: 982-1012.
113. Gasteiger J, Marsili M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. Tetrahedron 1980; 36: 3219-3228.
114. Chen R, Weng Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins 2002; 47: 281-294.
115. Chen R, Weng Z. A novel shape complementarity scoring function for protein-protein docking. Proteins 2003; 51: 397-408.
116. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003; 125: 1731-1737.

117. de Vries SJ, van Dijk AD, Krzeminski M, van Dijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AM. HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins 2007; 69: 726-733.

118. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. Integrating statistical pair potentials into protein complex prediction. Proteins 2007; 69: 511-520.

119. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997; 272: 106-120.

120. Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 2007; 67: 1078-1086.

121. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 1983; 4: 187–217.

122. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003; 331: 281-299.

123. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997; 267: 707-726.

124. Karaca E, Melquiond AS, de Vries SJ, Kastritis PL, Bonvin AM. Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server. Mol Cell Proteomics 2010; 9: 1784-1794.

125. Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W, Klein M L. Comparison of simple potential functions for simulating liquid water. J Chem Phys, 1983; 79: 926-935.

126. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE. Peptide folding: When simulation meets experiment. Angew Chem Int Ed 1999; 38: 236-240.

127. Jorgensen WL, Tirado-Rives J. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclin peptides and crambin. J Am Chem Soc. 1988; 110: 1657-1666.

128. Fernández-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 2004; 335: 843-865.

129. Méndez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 2003; 52: 51-67.

130. Eisenstein M, Katchalski-Katzir E. On proteins, grids, correlations, and docking. C R Biologies 2004; 327: 409-420.

131. de Vries SJ, Melquiond AS, Kastritis PL, Karaca E, Bordogna A, van Dijk M, Rodrigues JP, Bonvin AM. Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. Proteins 2010; 78: 3242-3249.

132. de Vries SJ, Bonvin AM. How proteins get in touch: Interface prediction in the study of biomolecular complexes. Curr Pept Prot Res 2008; 9: 394-406.

133. Lensink MF, Wodak SJ. Blind predictions of protein interfaces by docking calculations in CAPRI. Proteins 2010; 78: 3085–3095.

134. de Vries SJ, van Dijk M, Bonvin AM. The HADDOCK web server for data-driven biomolecular docking. Nat Protoc 2010; 5: 883-897.

135. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein-protein docking benchmark version 3.0. Proteins 2008; 73: 705-709.

136. Vajda S. Classification of protein complexes based on docking difficulty. Proteins 2005; 60: 176-180.

137. Rodriguez R, Chinea G, Lopez N, Pons T, Vriend G. Homology modeling, model and software evaluation: three related resources. Bioinformatics 1998; 14: 523-528.

138. de Vries SJ, Bonvin AM. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. Submitted

139. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. Protein-Protein Docking Benchmark 2.0: an update. Proteins 2005; 60: 214-216.

140. Kairys V, Gilson MK, Fernandes MX. Using protein homology models for structure-based studies: approaches to model refinement. ScientificWorldJournal 2006; 6: 1542-1554.

141. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins 2007; 69: 175-183.

142. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25: 3389-3402.

143. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000; 302: 205-217.

144. Heringa J. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. Comput Chem 1999; 23: 341-364.

145. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. Proteins 2009; 77: 18-28.

146. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH – a hierarchic classification of protein domain structures. Structure 1997; 5: 1093-1108.

147. Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A, Fanelli F, Gohlke H, Kuhn LA, Morris GM, Orozco M, Pertinhez TA, Rizzi M, Sotriffer CA. Target flexibility: an emerging consideration in drug discovery and design. J Med Chem 2008; 51: 6237-6255.

148. Wong CF. Flexible ligand-flexible protein docking in protein kinase systems. Biochim Biophys Acta 2008; 1784: 244-251.

149. Totrov M, Abagyan R. Flexible ligand docking to multiple receptor conformations: a practical alternative. Curr Opin Struct Biol 2008; 18: 178-184.

150. Carvalho AL, Dias FM, Prates JA, Nagy T, Gilbert HJ, Davies GJ, Ferreira LM, Romão MJ, Fontes CM. Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. Proc Natl Acad Sci U S A 2003; 100: 13809-13814.

151. Sansen S, De Ranter CJ, Gebruers K, Brijs K, Courtin CM, Delcour JA, Rabijns A. Structural basis for inhibition of Aspergillus niger xylanase by triticum aestivum xylanase inhibitor-I. J Biol Chem 2004; 279: 36022-36028.

152. Bonsor DA, Grishkovskaya I, Dodson EJ, Kleanthous C. Molecular mimicry enables competitive recruitment by a natively disordered protein. J Am Chem Soc 2007; 129: 4800-4807.

153. Meenan NA, Sharma A, Fleishman SJ, Macdonald CJ, Morel B, Boetzel R, Moore GR, Baker D, Kleanthous C. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proc Natl Acad Sci USA 2010; 107: 10080-10085.

154. Kim C, Lee B. Accuracy of structure-based sequence alignment of automatic methods. BMC Bioinformatics 2007; 8: 355.

155. Cavasotto CN, Kovacs JA, Abagyan RA. Representing receptor flexibility in ligand docking through relevant normal modes. J Am Chem Soc 2005; 127: 9632–9640.

156. Garzón JI, Kovacs J, Abagyan R, Chacón P. DFprot: a webtool for predicting local chain deformability. Bioinformatics 2007; 23: 901–902.

157. Dobbins SE, Lesk VI, Sternberg MJ. Insights into protein flexibility: the relationship between normal modes and conformational change upon protein-protein docking. Proc Natl Acad Sci USA 2008; 105: 10390–10395.

158. Keating KS, Flores SC, Gerstein MB, Kuhn LA. StoneHinge: hinge prediction by network analysis of individual protein structures. Protein Sci 2009; 18: 359-371.

159. Emekli U, Schneidman-Duhovny D, Wolfson HJ, Nussinov R, Haliloglu T. HingeProt: automated prediction of hinges in protein structures. Proteins 2008; 70: 1219-1227.

160. Schneidman-Duhovny D, Nussinov R, Wolfson HJ. Automatic prediction of protein interactions with large scale motion. Proteins 2007; 69: 764-773.

161. Shatsky M, Nussinov R, Wolfson HJ. FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. J Comput Biol 2004; 11: 83-106.

162. Hwang H, Vreven T, Pierce BG, Hung J, Weng Z. Performance of ZDOCK and ZRANK in CAPRI rounds 13–19. Proteins 2010; 78: 3104-3110.

163. Panjkovich A, Melo F, Marti-Renom MA. Evolutionary potentials: structure specific knowledge-based potentials exploiting the evolutionary record of sequence homologs. Genome Biol 2008; 9: R68.

164. Cavasotto CN, Orry AJW, Abagyan RA. The challenge of considering receptor flexibility in ligand docking and virtual screening. Curr Comput Aided Mol Des 2005; 1: 423-440.