Computational Analysis of Docked Protein-Protein Complexes

DISSERTATION

zur Erlangung des Doktorgrades der Naturwissenschaften

(Dr. rer. nat.)

der Fakultät für Physik

der Universität Regensburg

vorgelegt von

Florian Fink

aus Regensburg

im Mai 2011

Promotionsgesuch eingereicht am: 10.05.2011

Die Arbeit wurde angeleitet von: Prof. Dr. Wolfram Gronwald

Prüfungsausschuss:

Vorsitzender:	Prof. Dr. Josef Zweck
1. Gutachter:	Prof. Dr. Elmar Lang
2. Gutachter:	Prof. Dr. Ingo Morgenstern
3. Prüfer:	Prof. Dr. Thomas Niehaus

Abstract

The content of this work is, of course, condensed in the title. But what means "Computational Analysis of Docked Protein-Protein Complexes" in more detail? First of all, the objects of investigation are complexes between proteins. Not single proteins, not complexes between proteins and peptides and, to constrict it even more, only complexes of exactly two proteins, never more. So the analysis is done on dimeric protein complexes. Next, the title tells that the complexes are docked. That means, that the structures of interest did not arise out of experiments like X-ray crystallography or NMR spectroscopy but were calculated from docking algorithms. These algorithms take the experimentally solved structures from single proteins and simulate the process of complex formation. Their output is usually a huge number of putative complex conformations, which, in the best case, contains some near native structures. The native structure is the complex as it exists in nature. Near native structures have similar conformations as the native structure and are the optimum docking algorithms can reach. The big challenge is to find these near native structures among the - often more than 1000 - solutions. This subject was addressed during the here presented work by creating a scoring algorithm, which is able to judge the proposed solutions from docking algorithms. The developed PROtein COmplex analysis Server (PROCOS) is not only able to calculate a score for each solution and by

this provide a ranking that filters the best complexes to the top, as existing scoring algorithms do, but computes a probability for each complex to be native. This goal is achieved by calculating some energetically properties of a complex and compare these properties to those of a huge database of native and false complexes. Thereby, it is possible to decide to which group an investigated structure is more likely to belong: The native or the false complexes. The output of PROCOS is the probability that the analyzed complex belongs to the group of native complexes.

After developping PROCOS, the algorithm was extensively tested and compared to other scoring algorithms. Out of 96 native test complexes PROCOS identified 87 as near native (PROCOS-probability above 50%). Other algorithms always result in scores for the complexes. For this test case ZRANK obtained values between -814 and -14 and DFIRE between -234 and 301. In this simple example it becomes already clear that PROCOS is superior to other methods by means of the interpretation of the results. A probability gives an understandable information on a single structure. A score only helps to rank many results but does not state anything about the absolute qualities of the structures. Further tests on larger datasets showed that the performance of PROCOS to identify near native complexes is comparable to existing algorithms and in some cases even better.

In the last chapter two examples of docking applications are discussed that were performed during this work, too. This part addresses the step that has to be done before scoring: docking. In this context the docking program HADDOCK was used to take part in basic research on protein based drug development. The first study was done on the complex formation of Saratin, which can be extracted from the saliva of leeches, and Collagen, which is the main part of human tissue. This interaction is of special interest as it

was observed that Saratin prohibits blood coagulation and could therefore be used in a drug to prevent this mechanism. The docking experiment elucidated the complex formation of Saratin and Collagen, could identify the interface between the two proteins and predicted the conformation of the complex.
In the second study the melanoma inhibitory activity (MIA) protein was investigated. It is secreted from melanoma cells of skin cancer and causes the formation of metastases. Two docking experiments were done in this case: Since there is a hypothesis that MIA is only active as dimer, the complex structure of this dimer was modeled with HADDOCK. Then, in connection with the clinical research of finding a process to inhibit the formation of metastases formation, a putative complex formation of MIA and a small peptide AR71 was modeled. The fact, that the interface of the MIA dimer covers the same region as the peptide AR71 when it interacts with MIA, suggested to take AR71 into account as a deactivator of MIA. Further clinical investigations on mouse models actually showed a reduced formation of metastases on application of AR71.
In this work, the whole process of computer based prediction of protein complexes was studied with a strong focus on the last step of this process: The identification of near native protein complexes among 100s of putative docking solutions. The result is the scoring algorithm PROCOS, which is publically available on the internet under http://compdiag.uni-r.de/procos/.

Contents

1 Introduction **9**

1.1 Proteins 9

1.2 Protein-Protein Complexes 15

1.3 Docking 17

1.4 Motivation and Overview 19

2 HADDOCK **21**

3 PROCOS **25**

3.1 The Idea 25

3.2 From the Idea to the Program 28

3.2.1 Different Energies for Native and non Native Complexes 28

3.2.2 Datasets of Native and False Complexes 30

3.2.3 Three Scoring Functions for PROCOS 32

3.2.4 Electrostatic and van der Waals Scoring Functions . . . 33

3.2.5 Preprocessing 35

3.2.6 Calculation of Probabilities 37

3.2.7 Some Ideas for Combining the three Scores to a Single Probability 39

3.2.8 Using CAPRI Data as False Distributions 46

3.3 A General Overview 47
3.4 PROCOS in Detail 56
3.4.1 PHP-Scripts 56
3.4.2 Intermol 58
3.5 Testing PROCOS 61

4 Docking Applications **79**
4.1 Model of the Saratin-Collagen Complex 79
4.1.1 Background 80
4.1.2 Docking 80
4.2 MIA 83
4.2.1 Background 84
4.2.2 Docking 85

5 Summary and Outlook **91**

A CAPRI **95**

B Technical Remarks **99**
B.1 Intermol Without PROCOS 99
B.2 Creating Distribution Plots 100

Own Publications **101**

Bibliography **103**

List of Figures **110**

Chapter 1

Introduction

1.1 Proteins

Figure 1.1: Berzelius (1779 - 1848)

The Swede Jöns Jakob Berzelius is said to be the father of modern chemistry. To him not only the still common notation of chemical elements, the basic concepts of organic chemistry and the discovery of several elements can be traced back, but also did he give the proteins 1838 their name. The word is derived from the Greek word $\pi\rho\omega\tau\varepsilon\upsilon\omega$ (proteuo, "I take the first place", from $\pi\rho\omega\tau o\sigma$, protos, "the first", "the most important"). Therewith, Berzelius wanted to focus on the importance of proteins for life, which was already known at that time. And actually, proteins are not only decisively involved in the structural build-up of cells (as collagen) but also fulfill the major part of functions taking place in living creatures: As enzymes they control biochemical reactions in the body, as ion channels they regulate the ion concentration in the cell, as antibodies they serve infection

defense, as membrane receptors they recognize certain substances outside the cell and forward corresponding signals into the cell, etc. This list could be considerably extended and gives a good vision on the manifold roles that proteins play. For this reason they are often called the machines of the cell.

But how do proteins achieve such a variety of functions? The history of science has shown, that it is much easier for us to understand contexts and functionality in nature if we can see the object of examination. Therefor, X-ray studies have been used to determine the structure of chemical compounds since the beginning of the last century. The first x-ray crystallographic structural results on a globular protein molecule, myoglobin, reported in 1958 [1], came as a shock to those who had believed that they would reveal general simple principles of how proteins are folded and function, analogous to the simple and beautiful doublestranded DNA structure that had been determined five years before by James Watson and Francis Crick. John Kendrew at the Medical Research Council Laboratory of Molecular Biology, Cambridge, who determined the myoglobin structure to low resolution in 1958, expressed this disappointment in the following words: "Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicted by any theory of protein structure." [2] Today it is obvious that this complex structure is a precondition for proteins to fulfill their diverse functions.

As more protein structures were revealed, their shape was organized in a structural hierarchy to describe different molecules. Figure 1.4 explains the four levels graphically. Proteins are long chains of amino acids, linked by peptide bindings. The principle assembly of an amino acid is shown in Fig-

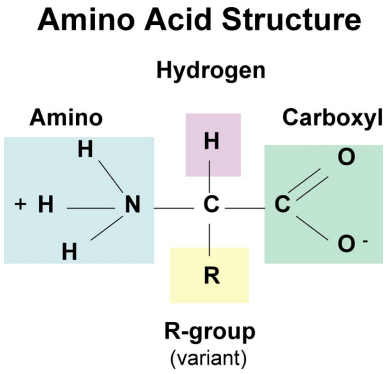


Figure 1.2: Principle structural assembly of amino acids. The R is called side chain and differs between the 20 natural amino acids.

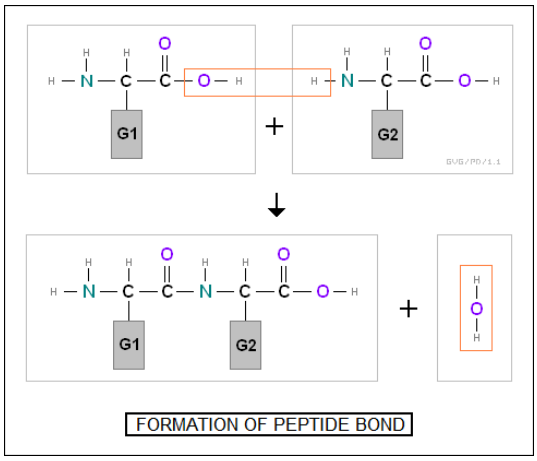


Figure 1.3: Two amino acids with side chains G1 and G2 are forming a peptide bond by precipitation of water.

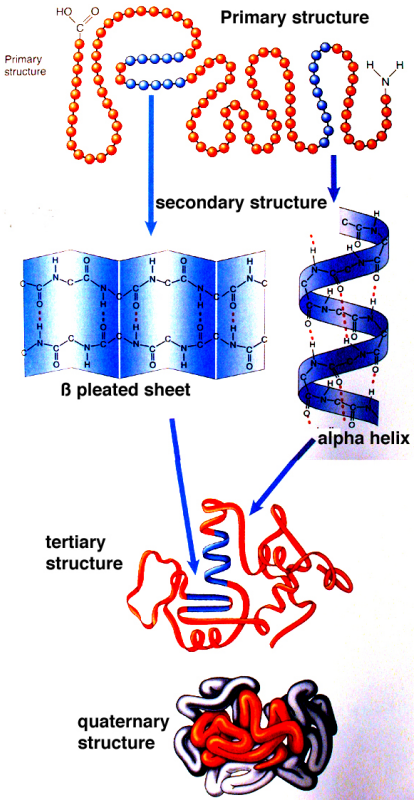


Figure 1.4: The primary structure is the sequence of the amino acids. Certain amino acid sequences form β sheets or α helicis to form intramolecular hydrogen bonds with the protein core. This is the secondary structure, which is stabilized by hydrogen bonds, marked with red dotted lines in the figure. The spatial arrangement of the β sheets and α helices is called tertiary structure. Bigger proteins consist of several domains and the arrangement of these domains is called quaternary structure.

ure 1.2. It consists of a central C-atom, which is bound to a hydrogen, a carboxyl- and a amino group and a side chain R, which individualizes the 20 natural existing amino acids. The process of peptide binding that connects the amino acids together to form the proteins is shown in Figure 1.3. One proton from the amino group and a hydroxide from the carboxyl form a water molecule and the two amino acids are connected together. The order of the 20 natural existing amino acids in these chains is coded in the genes on the corresponding DNA-strand. In this context, one amino acid is unambiguously defined by three consecutive base pairs in the DNA. From the DNA messenger RNA (mRNA) is built, that codes for a specific protein. The mRNA is transfered to the ribosomes where it is used as a blueprint for protein production according to the just mentioned "translation code". The order of the amino acids in the complete protein is called its sequence or its primary structure. The chains do however not remain in an outstretched state, but try to assume an energetically favorable shape. This process is called protein folding and it can be observed, that identical sequences always fold in the same way. The mechanisms, which drive the folding are still a subject of present research but it is clear that the main contribution to the process is a minimization of the Gibbs free energy $\Delta G = \Delta H - T\Delta S$, where H is the enthalpy, T the temperature and S the entropy of the system. This behavior expresses itself by the tendency of proteins to fold in a way that hydrophobic amino acids avoid contact with the surrounding water and two types of secondary structure are formated inside the protein: alpha helices and beta sheets. The formation of secondary structure is to some extend determined by the primary structure. Certain amino acid sequences favor either α helices or β strands. Secondary structure elements usually arrange themselves in simple motifs, by packing side chains from adjacent α helices

or β strands close to each other. Several motifs usually combine to form compact globular structures, which are called domains. This is the tertiary structure of proteins. Bigger proteins can have more than one domain. In this case the spatial arrangement of the domains is called quarternary structure. However, the quarternary structure does not affect the folding of the domains. In other words a domain would fold in the same way if separated from the rest of the protein.

To understand the functions of proteins in detail it is important to know their 3D structure. As mentioned above the first and still most common method to determine the structure of proteins is x-ray crystallography. This method allows one to measure the 3D density distribution of electrons in the crystallized state of proteins. To date, 86.8% of the 69510 entries in the PDB archive [3] are solved by this method. Only in the mid 1980s another competing method came in use: Nuclear Magnetic Resonance (NMR) spectroscopy. The great advantage of this new method was the possibility to analyze the proteins close to their physiological conditions. However, NMR spectroscopy is more limited with regard to the protein's size. 12.5% of the PDB entries are derived from NMR. The remaining 0.7% (only 502 structures) were determined by other methods like electron microscopy. A comprehensive explanation of the two most common methods can be found in the last chapter of the book of Branden and Tooze [2].

The analysis of a protein sequence is much simpler than the determination of a protein structure. However, the structure of a protein gives much more insight in the function of the protein than its sequence. Therefore, a number of methods for the computational prediction of protein structure from its sequence have been proposed. The challenge to model a protein structure from its sequence by only using physical interactions as driving forces could

only be solved for small proteins so far. This is mostly due to the vast computational resources these so called *ab initio*- or *de novo*- methods require. Comparative modeling on the other hand is already very effective. In this context, known structures from proteins with similar sequences are used as starting point for the structure prediction. Homology modeling and protein threading are the two approaches that use this trick to reduce the required computational resources to a reasonable amount. The recent progress and challenges in protein structure prediction are reviewed by Zhang [4].

1.2 Protein-Protein Complexes

Knowledge about single proteins is good but not enough to understand what is happening in a cell. Actually, most processes in the cell are carried out by complexes of proteins. For example inhibitor proteins that deactivate enzymes by connecting to their active centers (see Figure 1.5), DNA binding proteins that need to be in contact with an activating domain to start the transcription, the ribosome, which performs the translation from RNA to protein sequences (see Figure 1.6), membrane receptor proteins that are waiting for certain proteins to connect to, which will initialize some process on the other side of the membrane, etc. It has been estimated that each protein has on average nine interaction partners [5].

Protein complexes are a form of quaternary structure. The physical motivation for proteins to form complexes is quite similar to the effects that drive protein folding. If the free energy is reduced by two proteins when they come near to each other in a certain orientation and their surface shapes admit a proximal contact it is likely that they form a complex. During the process of complex formation it is possible that the involved proteins un-

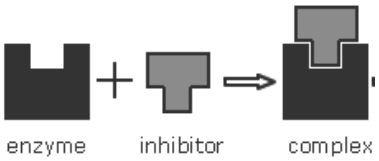


Figure 1.5: Schematic demonstration of a enzyme-inhibitor complex. Normally, on complex formation, some function of the enzyme is inhibited.

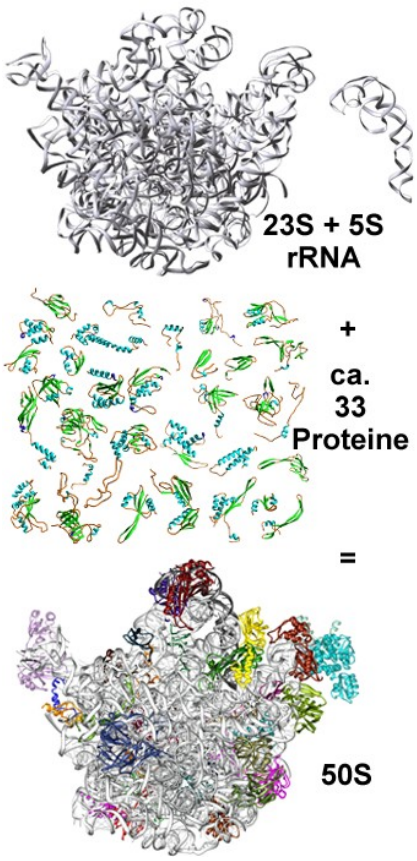


Figure 1.6: A good example for a very big protein complex is the ribosome. It comprises a small (30S) and a big (50S) subunit of which only the 50S-unit is shown. Even this part consists of about 33 single proteins and several rRNA strands.

dergo conformational changes to adopt in the energetically best way to each other. This is the reason why structural information about single proteins is normally not clarifying the structure of the complex they build. On the other hand it is not always possible to determine the structures of protein complexes by experimental methods due to limitations concerning large or transient complexes. In addition the experimental structure determination of protein-protein complexes is in most cases a very time-consuming and challenging process. However, there are many well established methods to detect protein-protein interactions, like yeast2hybrid assays [6, 7] or tandem-affinity-purification mass spectrometry [8]. These experimental approaches are supplemented by bioinformatic methods such as phylogenetic profiling [9], investigations of gene neighborhoods, and gene fusion analysis. Unfortunately these methods only give information on, which proteins interact and say nothing about the spatial structure of the complexes. For that reason computational approaches like docking algorithms that predict the structure of these complexes are needed.

1.3 Docking

Docking is the computational prediction of protein complex structures from the unbound structures of the single proteins. It is normally performed in three steps:

(i) Sampling

A huge number of randomized start confirmations of the protein structures to be docked is created.

(ii) Optimizing

The start complexes are optimized with respect to different energy terms,

geometrical reasons or other supplemental information by performing translational, rotational and conformational changes.

(iii) Scoring

The resulting structures are ranked to find the best solutions (that is to say those that are most similar to the native complex structure) in top positions. This is achieved by calculating appropriate scores that measure the quality of the complexes.

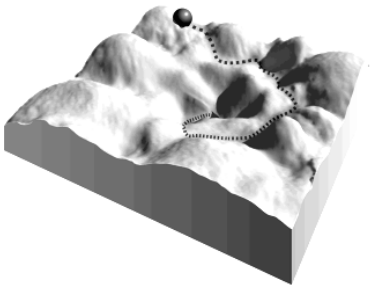


Figure 1.7: Each point in this energy landscape represents one conformation of the complex. Docking algorithms have to find a way (dotted line) from the randomized start confirmation (sphere) to the global minimum

Docking approaches assume that the native complex is near the global minimum of the energy landscape constituted by the set of all theoretically possible complex conformations of the interacting proteins. The main challenge of any algorithm is to find this minimum (see Figure 1.7). Since the size of this landscape is immense, the sampling step is very important to make sure that at least some start conformations are sufficiently close to the global minimum to find it during optimization. The great number of optimized structures on the other hand creates the problem of choosing the best solutions in the end. This is tackled by the scoring step. Usually several factors are considered in the identification of near native models. These include steric surface

complementarity [10], electrostatic interactions [11], hydrogen bonding [12], knowledge based pair-potentials [13], desolvation energies [14] and van der Waals interactions [15]. It has been shown that scoring can be improved considerably by combining the information of several scoring functions [16], and this is increasingly becoming common practice [17, 18, 19].

During the last decade considerable effort has been put in the development and application of docking algorithms; for a review see [20]. The success of docking algorithms has consistently improved over the last years as measured by the CAPRI blind docking experiment [21]. Due to such efforts, not only is the reliability of *in silico* docked complexes becoming more widely accepted but the various available docking algorithms can be objectively compared. In spite of many successful developments in this area it still remains a lot of work to be done in the challenge of docking. Lensink et al. stated in a recent overview of the results of the CAPRI experiment, that large conformational adjustments are still not handled satisfactorily and that scoring methods are not sensitive enough to identify the best models [22].

1.4 Motivation and Overview

During the work presented here the current docking and scoring problems were analyzed and tested and especially the scoring landscape was expanded by a scoring algorithm based on a new scoring method. In this context, the docking algorithm HADDOCK, which is presented in the following chapter, was used to model the 3D structure of biological relevant protein complexes. This was mostly done in collaboration with biological and medical institutes and the two main projects are discussed in chapter 4. The main part of this work was, however, dedicated to the development of the just mentioned

novel scoring algorithm for docked complexes: PROCOS. Chapter 3 explains in detail the whole process from the idea to the functionality of the finished program.

Chapter 2

A Dockingprogram used in this Context: HADDOCK

All docking runs that were performed in the course of this work made use of the docking program HADDOCK, version 2.0 [19, 23]. HADDOCK is an up to date docking algorithm that allows the user to add supplemental knowledge about binding sites by means of so called ambiguous interaction restraints (AIRs) and provides the possibility to account for conformational flexibility of side chains and backbone.

An AIR is defined as an ambiguous intermolecular distance d_{iAB} between active and passive residues of the proteins A and B (see Figure 2.1). They are incorporated in the optimization process as an additional energy term that has to be minimized. Residues that are defined as active by the user have to be part of the binding site, passive residues may be part of it. During docking an effective distance

$$d_{iAB}^{eff} = \left(\sum_{m_{iA}=1}^{N_{atoms}} \sum_{k=1}^{N_{res}B} \sum_{n_{kB}=1}^{N_{atoms}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-\frac{1}{6}} \tag{2.1}$$

between any atom m of an active residue i of protein A (m_{iA}) and any atom

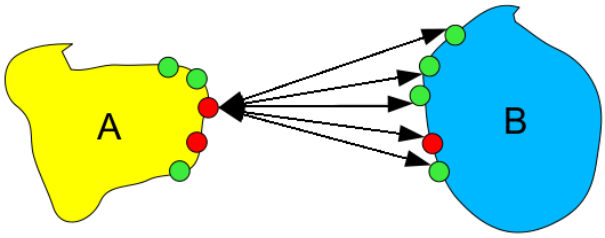


Figure 2.1: Symbolic visualization of the concept of AIRs. Active residues are marked in red, passive residues in green. To fulfill the restraints, at least one distance of every active residue to all active and passive residues of the partner protein has to be less then 2 Å.

n of both active and passive residues k of protein B (n_{kB}) is calculated. N_{atoms} indicates the number of all atoms in a given residue and N_{res} the sum of active and passive residues for a given protein. In this way, the passive residues do not have direct AIRs to the partner protein but can satisfy the partner proteins active restraints. To fulfill the restraint, d_{iAB}^{eff} has to be smaller then 2 Å.

The docking protocol is performed in three stages:

(i) Randomization of orientations and rigid body energy minimization.

The two proteins are positioned at 150 Å from each other in space and each protein is randomly rotated around its center of mass. Then the proteins are allowed to rotate to minimize the intermolecular energy function. Afterwards both translations and rotations are allowed, and the two proteins are docked by rigid body energy minimization. The best solutions in terms of intermolecular energies are then further refined in the next step.

(ii) Semirigid simulated annealing in torsion angel space.

The second stage consists of three simulated annealing refinements at different temperature ranges. First the two proteins are considered as rigid bodies and their respective orientation is optimized. Then the side chains

at the interface are allowed to move and in the third step both side chains and backbone at the interface are allowed to move to allow for some conformational rearrangements.

(iii) Final refinement in Cartesian space with explicit solvent.

The final stage consists of a refinement in an 8 Å shell of TIP3P water molecules. This is a model of the water molecule, often used in computational chemistry to approximate molecular mechanisms. In the TIP3P model each atom gets assigned a point charge, and the oxygen atom also gets the Lennard-Jones parameters. The model uses a rigid geometry matching the known HOH angle of 104.5°. More details about this water model can be found at Jorgensen et al. [24].

Although no real significant structural changes occur during the water refinement stage, it is useful for the improvement of the energies, which is important for a proper scoring of the resulting conformations. To calculate the score for the ranking, HADDOCK summes up desolvation energy (1.0), intermolecular electrostatic energy (0.2), intermolecular van der Waals energy (1.0) and violation of AIRs (0.1), weighted by the factors given in brackets.

HADDOCK has participated in the CAPRI experiment since round 4 and has shown excellent prediction and scoring results in comparison to other groups in the last years.

Chapter 3

PROCOS

The PROtein COmplex analysis Server PROCOS [25] is a webserver based scoring algorithm, which admits the user to upload his pdb-files of dimeric protein complexes that were obtained by docking algorithms or any other method. The program then calculates the probability for these complexes to be native. PROCOS was developed from the ground during this work and the method is ready to use under http://compdiag.uni-r.de/procos/. Figure 3.1 shows the starting page of PROCOS.

3.1 The Idea

As mentioned in the first chapter, scoring of docked protein complexes is a challenging task. Current scoring algorithms are still not able to reliably identify near native structures [22]. Therefore, it is desired to develop more sophisticated methods.

The goal for PROCOS was to develop an easy to use scoring algorithm that produces intuitively interpretable outputs and to compare its results to existing programs. Scoring is, as in HADDOCK, often done with a pseudo-energy

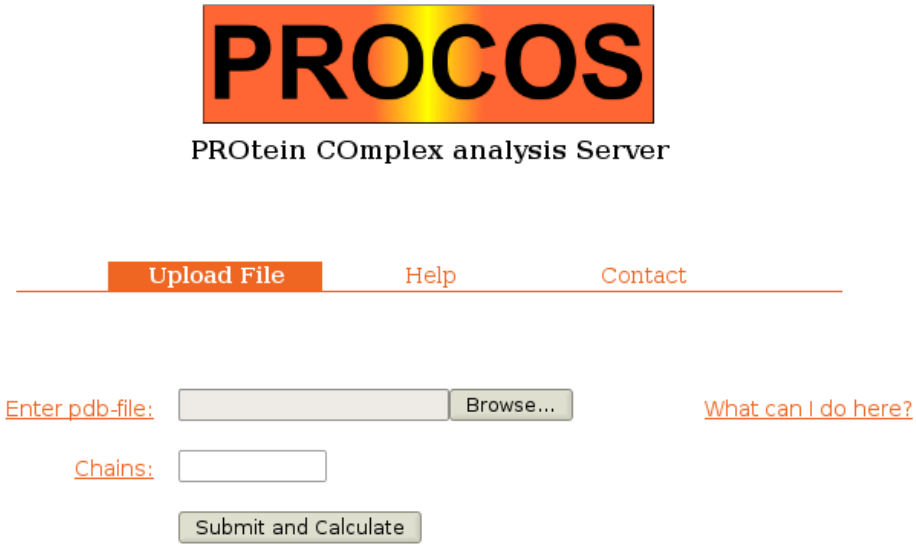


Figure 3.1: Screenshot of the PROCOS home page.

term that is a weighted combination of real energies and other factors like buried surface area or knowledge based pair-potentials. Such a score allows to sort docking solution. At best, eventually existing near native complexes among the structures are sorted to the top of such a list. However, even ensembles with no near native solutions will be sorted in some way and as the score is only a number that allows to compare the different solutions, no assertion about the quality of the top ranked complexes is given.

The "score" PROCOS calculates is the probability that the assigned structure represents a native complex. Thereby, even the top ranked complexes may get a low probability to be native and it is possible to observe that no usable results were produced during docking. To obtain such a probability it is necessary to compare a questioned complex to a set of native and not existing (false) complexes and to find a measure that decides with which probability the complex belongs to either of these groups. In the preliminary work to PROCOS it could be shown that score values obtained from amino

acid based pair-potentials distributed differently for a set of native complexes compared to a set of false complexes [26]. In Figure 3.2 it can be seen that scores from the two groups are not totally separated, but that there is an evident difference in the shape and the position of the two curves. This difference is the bases of PROCOS' ability to assign a complex with a certain probability to one of these two groups.

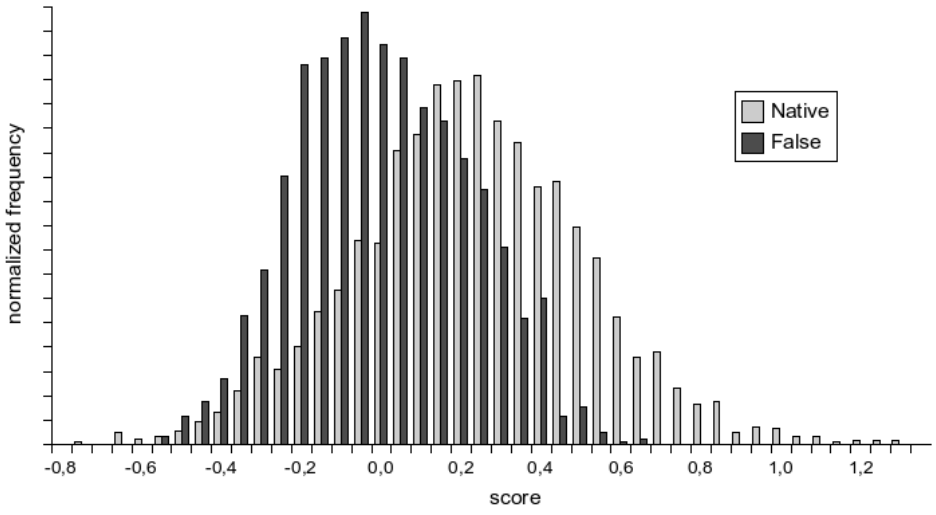


Figure 3.2: Score distributions from amino acid based pair-potentials for native and false complexes.

Murphy et al. have shown that scoring can be improved considerably by combining the information of several scoring functions [16]. PROCOS, therefore, combines intermolecular electrostatic energy, van der Waals energy and an amino acid based pair-potential in its probability calculation. The program is implemented in a modular architecture, which makes it easy to include further scores in future.

In the sequel of this chapter PROCOS will be presented. The next section summarizes the process of developping PROCOS. Then an overview of the current functionality of PROCOS is given. Section 3.4 explains the

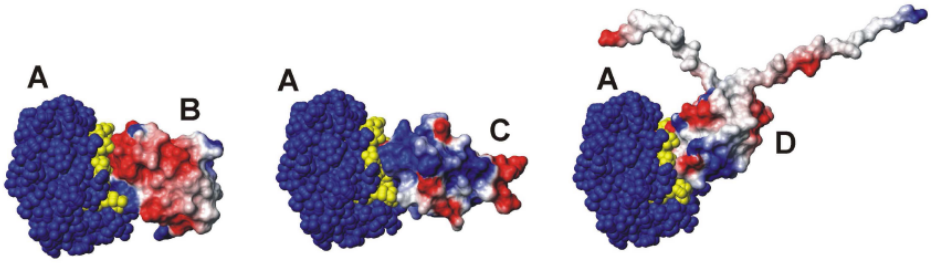
program structure in detail and gives all information that is necessary to further develop the algorithm. Finally, the last section in this chapter will present results that were obtained with PROCOS and shows its performance in comparison to other scoring algorithms.

3.2 From the Idea to the Program

3.2.1 Different Energies for Native and non Native Complexes

To give PROCOS a chance to work it is a precondition to find energy terms that have lower (better) values for native protein-protein complexes than for non native complexes. Native protein-protein complexes define in this regard the interaction of proteins that interact in nature and non native complexes are formed by proteins that do not interact in nature. To verify this requirement, several proteins were docked. To find appropriate proteins the benchmark sets from Mintseris et al. and Hwang et al. [27, 28] were used. Appropriate in this context means, that the structures of the single proteins are solved as well as the complex structure and that they are accessible in the PDB. The single proteins are used to dock the complex and the native complex is needed to verify the quality of the docking and to test if the ranking was meaningful. To make the docking realistic, the single protein structures were taken in their unbound state so that the docking algorithm had to model eventually occurring conformational changes. In the first attempt the facility of HADDOCK to define the interacting residues was not used since it should be investigated whether the correct orientation of the molecules can be found by the algorithm without additional data. Unfor-

tunately, this approach was unsuccessful, which means that no near native structures were found by HADDOCK. Thus, a second test with slightly more information was performed where one side of the interface was defined by active residues, whereas the interface side of the smaller docking partner was still undefined. This was achieved by defining all surface amino acids of this molecule as passive. This concept was used for all docking runs performed in the sequel where nothing else is mentioned. Figure 3.3 shows the results for



Interaction A-B	Interaction A-C	Interaction A-D
E_{inter} -2512,2 kJ/mol	E_{inter} -1991,4 kJ/mol	E_{inter} -1695,6 kJ/mol

Figure 3.3: Average interactions energies for three different trial molecules (B, C, and D) docked to the cytoplasmatic A domain (IIA(MTL)) of the mannitol transporter II (A). (B) histidine containing phosphocarrier protein (HPr), (C) human cyclin dependent kinase subunit type I (CKSHS1), and (D) apo form of HMA domain of copper chaperone for superoxide dismutase.

the very first tests. Three different molecules were docked to the cytoplasmatic A domain (IIA(MTL)) of the mannitol transporter II, marked A and shown in blue. The defined interface is colored yellow. The histidine containing phosphocarrier protein (HPr), marked B, forms the native complex with A. C and D do not build complexes with A in nature but were forced by docking to do so. For each complex the average interaction energy (sum of electrostatic and van der Waals energy) of all 200 solutions is given. The

average energy of the near native complexes is clearly lower (better) then the other energies, which gives a first hint on the feasibility of PROCOS. In principle these investigations show that by the use of docking calculations it is not only possible to obtain the correct 3D structure of a protein-protein complex, but that it is also possible to discriminate between proteins that do interact in nature and those that do not interact.

In the following studies more complexes of docked proteins were analyzed in the same manner [29]. Table 3.1 shows results of these tests. The native complexes, which are shaded in yellow obtain not always the lowest energy, but the trend is confirmed. Note that in this case the shown energy is only the mean of the top 10 ranked solutions of HADDOCK and not of all 200 solutions. This was done to avoid exploitation of badly docked complexes.

3.2.2 Datasets of Native and False Complexes

The just described preliminary investigations confirmed the principal possibility to discriminate between native and non native complexes. To put the analysis on a more stable basis, two datasets were needed: Native complexes and false complexes. The native complexes were taken from the Mintz database [30]. It contains 2541 experimentally solved, non homologous native protein-protein complexes. This dataset is called Ndat in the following. Since a database with false complexes does not exist it had to be artificially created by a docking routine. For creating false complexes one cannot simply join two proteins in an arbitrary way since the resulting complexes would be extremely unrealistic. For a more realistic test set, false complexes are needed that do not exist in nature but are, nevertheless, optimized in a way that they could theoretically exist. This problem was tackled as follows:

The more than 5000 proteins constituting the complexes of Ndat were paired

Receptor	**Ligand**	E_{inter} **[kJ/mol]**
Barnase	Barstar	-913.2
Barnase	Soybean trypsin inhibitor	-670.0
Barnase	Ovomucoid 3rd domain	-575.0
Barnase	Eglin C	-510.6
Barnase	Pancreatic secretory trypsin inhibitor	-504.7
Barnase	APPI	-481.3
α-Chymotrypsin	Eglin C	-552.8
α-Chymotrypsin	Barstar	-505.5
α-Chymotrypsin	APPI	-445.7
α-Chymotrypsin	Soybean trypsin inhibitor	-364.9
α-Chymotrypsin	Pancreatic secretory trypsin inhibitor	-306.3
Bovine trypsin	CMTI-1 squash inhibitor	-588.4
Bovine trypsin	Glycosylase inhibitor	-761.3
Bovine trypsin	RAGI inhibitor	-492.2
Bovine trypsin	Soybean trypsin inhibitor	-436.5
Bovine trypsin	Streptomyces subtilisin inhibitor	-412.6
Bovine trypsin	Amicyanin	-323.9

Table 3.1: Comparison of intermolecular interaction energies of native (shaded in yellow) and corresponding non native complexes. The energy is always the average of ten complexes that were top ranked from the docking algorithm.

by chance. At the surface of all proteins interface areas of similar size were assigned at random. For this purpose the function *ranair*, which is part of HADDOCK was used. Utilizing these randomly chosen protein pairs, docking runs were performed. The top ranked complex of each such docking run was then incorporated into a dataset of false complexes. Thus, a dataset of reasonable non native complexes was produced, where each complex passed through a docking procedure with energy minimization and local structure improvement. This represents a meaningful antipode to the group of the native complexes. In total the group of false complexes contained 2440 members. This dataset is called Fdat1, as there will be another, better dataset of false complexes introduced later on.

3.2.3 Three Scoring Functions for PROCOS

In the next step appropriate properties of the complexes had to be chosen that can discriminate between native and false complexes. As in the preliminary tests, the intermolecular electrostatic and van der Waals energies were chosen. In addition an amino acid based pair-potential that was recently derived from the work of Wolowski [31] came in use. A pair-potential is a knowledge based scoring function, that deduces from a database of experimentally solved complexes the frequency that certain atom types or amino acids are part of the interface of protein complexes. In this case, Wolowski used Ndat as basis for the analysis as well and calculated scores for each amino acid pair that reflect the frequency to find this pair in the interface according to the following formula:

$$S_{inter}\left(aa_1, aa_2\right) = log\left[\frac{f_{pair}\left(aa_1, aa_2\right)}{f_{surface}\left(aa_1\right) f_{surface}\left(aa_2\right)}\right] \tag{3.1}$$

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.543																			
Cys	0.443	3.657																		
Asp	-0.818	-0.988	-0.846																	
Glu	-0.939	-0.837	-1.932	-0.925																
Phe	0.901	1.719	-0.080	-0.092	2.955															
Gly	-0.711	0.337	-1.013	-1.280	0.510	-0.457														
His	-0.108	0.145	0.024	0.179	1.054	-0.297	1.559													
Ile	0.640	0.928	-0.675	-0.587	1.730	0.167	0.545	1.928												
Lys	-0.902	-0.390	-0.375	-0.363	0.107	-0.940	-0.809	-0.146	-1.042											
Leu	0.377	0.521	-0.640	-0.146	1.905	0.114	0.143	1.539	-0.480	1.848										
Met	0.894	0.837	-0.812	-0.150	1.713	0.176	0.650	1.427	-0.214	1.576	2.877									
Asn	-0.322	0.078	-0.736	-0.876	0.656	-0.303	0.122	-0.060	-0.611	0.093	0.432	0.390								
Pro	-0.222	0.389	-0.727	-0.812	1.119	-0.621	-0.128	0.330	-0.846	0.383	0.640	-0.204	0.390							
Gln	-0.469	-0.120	-0.817	-0.831	0.821	-0.493	-0.171	-0.105	-0.549	0.427	0.611	-0.076	-0.335	0.354						
Arg	-0.284	0.629	0.332	0.242	0.660	-0.299	-0.274	0.433	-0.995	0.426	0.630	-0.35	0.074	0.130	0.639					
Ser	-0.635	0.173	-0.853	-0.726	0.515	-0.804	-0.199	0.329	-0.853	-0.068	0.256	-0.479	-0.405	-0.366	-0.366	-0.015				
Thr	-0.206	0.654	-0.923	-0.721	0.679	-0.671	0.059	0.646	-0.669	0.116	0.251	-0.481	-0.265	-0.168	-0.197	-0.345	0.257			
Val	0.304	1.041	-0.512	-0.326	1.585	-0.439	0.402	1.117	-0.463	1.286	1.537	-0.112	0.420	0.075	0.401	0.215	0.376	1.499		
Trp	0.691	1.644	0.225	0.205	1.952	0.434	1.164	1.574	0.161	1.674	1.163	0.810	0.946	0.106	0.501	0.311	0.755	1.355	3.052	
Tyr	0.280	0.968	0.012	0.375	1.685	0.038	1.064	1.376	0.189	1.357	1.289	0.385	0.775	0.406	0.669	0.268	0.332	1.132	1.376	1.778

Figure 3.4: Values for the pair-wise potentials found from Wolowski [31].

Here, $f_{pair}(aa_1, aa_2)$ is the frequency of finding two amino acids from different proteins separated by less then 0.5 Å between their closest van der Waals surfaces. Whereas, $f_{surface}$ is the frequency of a given amino acid being on the surface of a protein. Both, f_{pair} and $f_{surface}$ were calculated from all members constituting Ndat. Using Equation 3.1, a positive score means that it is likely to find a certain pair in the interface whereas a negative score means that such a pairing is unlikely. The resulting values of Wolowskis work are shown in the table in Figure 3.4. The term "pair-potential of a complex", which will be used in the further course of this work, is simply the sum of all individual scores of amino acid pairs that were found to be nearer to each other than 0.5 Å. The values were read out of Figure 3.4.

3.2.4 Electrostatic and van der Waals Scoring Functions

As electrostatic interaction between the surface atoms is probably the major force that drives the proteins into their native complex conformation it should be part of a good scoring algorithm. Compared to electrostatics, the

values of the van der Waals interaction is quite small, normally a factor 10 below the electrostatic interaction. However, van der Waals interaction becomes important for the fine-tuning of the structure. Atoms that attract each other due to opposite charges would come arbitrarily near to each other in a simulation without other forces and produce severe atom clashes. Van der Waals forces inhibit this behavior as they include the Pauli repulsion that reaches extremely high positive values as soon as the electron orbits of the atoms get in contact with each other. Therefore, both, electrostatic and van der Waals interaction were included as scoring functions into the PROCOS prediction. The exact model of the two forces used in PROCOS is similar to that used in CNS, which is used for the HADDOCK algorithm:

The electrostatic energy is the sum of the individual electrostatic energies of all intermolecular atom pairs in the complex. It is calculated according to the following equation:

$$E_{elec} = \sum_{n,m} \frac{q_n q_m C}{\varepsilon_0 R} \left[1 - \frac{R^2}{R_{off}^2} \right] \tag{3.2}$$

where n and m enumerate the atoms of the first and second protein, respectively; q is the charge of an atom; C is a scaling factor (set to 900); ε_0 the dielectric constant is set to one, as it is difficult to determine or approximate an exact value for ε_0 inside the very inhomogeneous matter of proteins; R denotes the distance between the atoms. The term in brackets ensures that the electrostatic energy approaches zero at a cut-off value of $R_{off} = 8.5$ Å. This cutoff saves computation time and the introduced error is negligible.

The van der Waals energy is a combination of the Pauli repulsion and the van der Waals attraction. This interaction between uncharged and not chemically bound atoms is in physical chemistry mostly modeled with a Lennard-Jones-Potential, which is a special case of the Mie Potential [32]. Here, the

Lennard-Jones-(12,6)-Potential is used, which means that the repulsive part of the equation is modeled with the 12th power. The van der Waals score is then calculated similar to the electrostatic energy as a sum of the Lennard-Jones-(12,6)-Potential over all intermolecular atom pairs, using the following equation:

$$E_{vdw} = \sum_{n,m} 4\varepsilon \left[\left(\frac{\sigma}{R} \right)^{12} - \left(\frac{\sigma}{R} \right)^{6} \right] SW\left(R, R_{on}, R_{off}\right) \tag{3.3}$$

with

$$SW = \begin{cases} 0 & \text{if } R > R_{off} \\ \frac{\left(R^2 - R_{off}^2\right)^2 \cdot \left(R^2 - r_{off}^2 - 3\left(R^2 - R_{on}^2\right)\right)}{\left(R_{off}^2 - R_{on}^2\right)^3} & \text{if } R_{off} > R > R_{on} \\ 1 & \text{if } R < R_{on} \end{cases} \tag{3.4}$$

where ε and σ parameterize the Lennard-Jones potential of identical atom types. Between different atom types, the following combination rule is used: $\sigma_{ij} = \frac{\sigma_{ii}+\sigma_{jj}}{2}$ and $\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}}$. The individual values are similar to those used by HADDOCK2.0 (see HADDOCK distribution, file "toppar/parallhdg5.3.pro" line 1095 ff) [19, 23]. R_{on} and R_{off} were set to 6.5 Å and 8.5 Å, respectively. Figure 3.5 shows a comparison of the electrostatic and van der Waals values between HADDOCK and the scoring functions used in this work. They are obviously not the same, as different program structures, cut-offs and parameter values are used, but it is clear that they have the same trend so that one can assume that the same physical interaction is measured.

3.2.5 Preprocessing

In order to generate more reliable predictions, we were interested in combining the different property functions. As the above functions are very different in their physical meaning, rescaling of the individual functions is

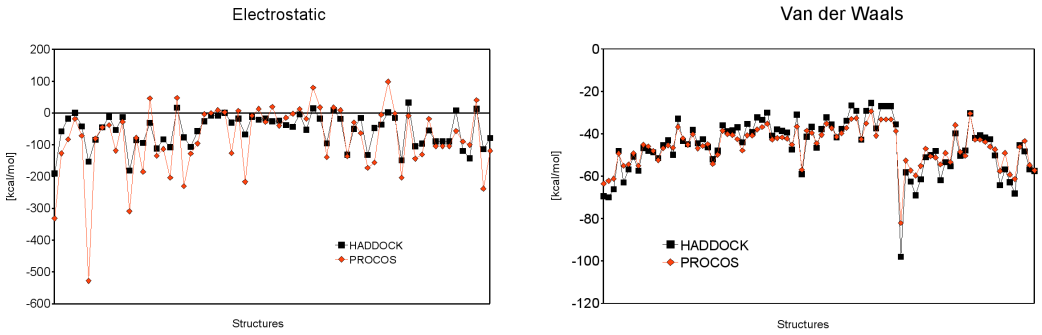


Figure 3.5: Comparison of the values of electrostatic and van der Waals energy from HADDOCK and PROCOS. The values are not identically but showing the same trend, indicating that the same physical behavior is measured.

required prior to their combination. Therefore, all data were rescaled to values between 0 and 1000, where 0 means worst and 1000 means best. In a first attempt the conversion factores that were used for this rescaling were defined manually by looking at the scores of Ndat and making sure that only very few extreme complexes obtained values below 0 or above 1000. Note that later a more precise method was used for rescaling. Table 3.2 shows the values for the rescaling, which can be used to rescale arbitrary values using the following equation:

$$new\_value = \frac{old\_value - rescaled_0}{rescaled_{1000} - rescaled_0} \cdot 1000 \tag{3.5}$$

For some complexes of Ndat extremely high E_{vdw} values resulted. A visual inspection of these complexes showed the existence of severe atom clashes. Since native complexes should ideally not show extreme clashes, all native complexes having a higher E_{vdw} value than the worst false complex were excluded from further analysis. By this action, 310 structures from Ndat were removed remaining 2231 structures to represent the native complexes. This reduced dataset is called Ndat-300.

	elec	vdw	pair
$rescaled_0$	83.55 kcal/mol	64.62 kcal/mol	-21.27
$rescaled_{1000}$	-2627.07 kcal/mol	-513.66 kcal/mol	233.67

Table 3.2: Conversion factors for the rescaling of the scoring functions. The table shows the original values for the rescaled scores of 0 and 1000. Inserting these values in Equation 3.5 converts an arbitrary value. Note that the pair potential has no unit.

3.2.6 Calculation of Probabilities

At this point, for every complex from the class of native and false complexes three scores were calculated. To visualize this, one could plot the results for the native and false complexes in different colors, as shown in Figure 3.6 for the electrostatic energy. The diagram shows clearly, that the scores of native complexes are differently distributed than the scores of false complexes. However, there is a better way to plot this data, which is easier to interpret and opens the possibility to assign other, unknown complexes to either of the two classes. This is to plot probability densities. The probability density defines for every score interval a probability to find a complex with a score within this interval. To obtain such a distribution of probability densities from the data, the following method was applied:

From every data-point n and its m neighbors the mean μ_n and the variance σ_n were calculated. These values were used to derive a gauss function for the corresponding data-point. In the end, the Gaussians for all data-points were added to produce the density D. The formula for this density is

$$D(x) = \sum_n \frac{1}{\sigma_n \sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x-\mu_n}{\sigma_n}\right)^2\right] \tag{3.6}$$

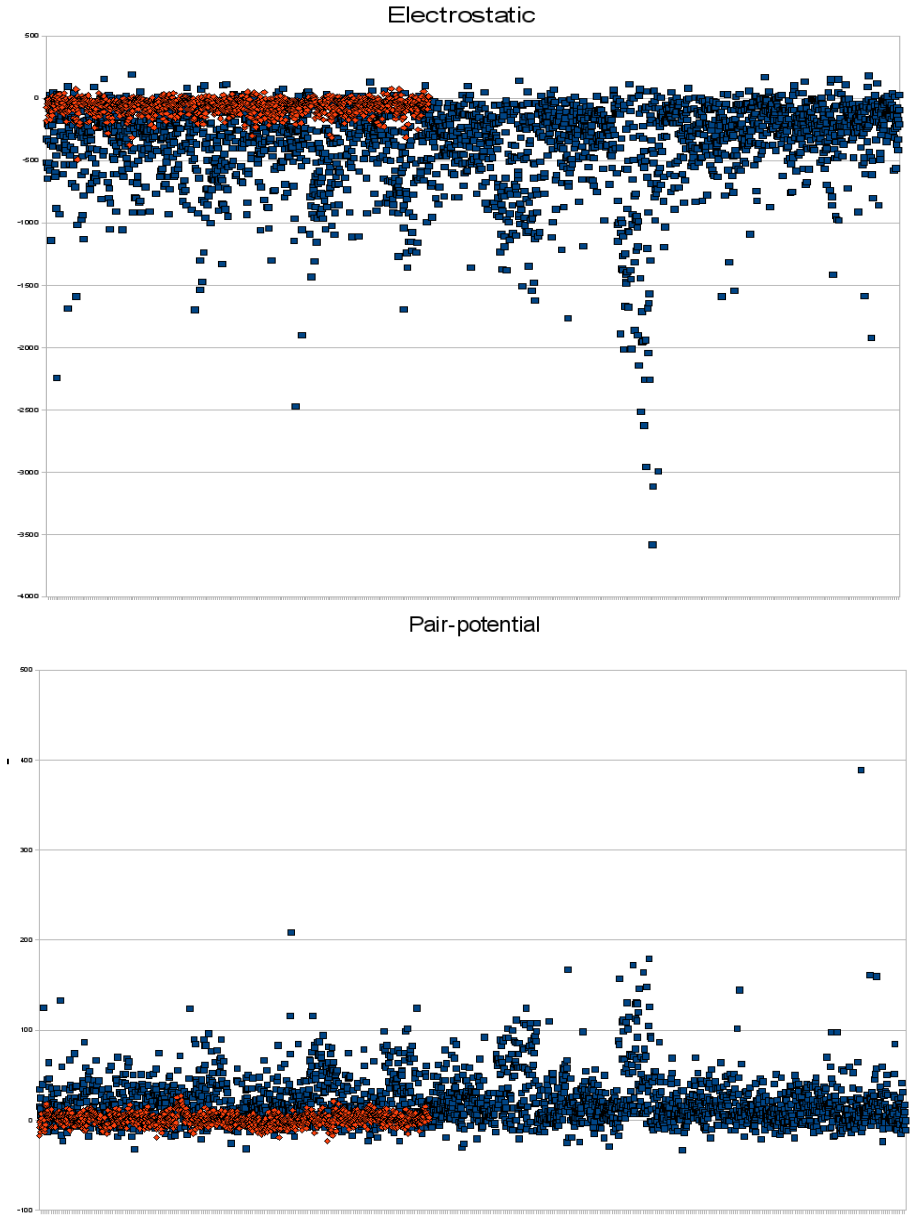


Figure 3.6: Electrostatic energy in kcal/mol and pair potential for all complexes of Ndat (blue) and Fdat1 (red). The different distribution of values for the two groups is obvious but the representation is not useful.

The parameter m (number of neighbors) determines the degree of smoothing and was for the first approaches set to 100. This value was set to ensure that the resulting curves should neither have to many peaks which would overemphasize single structures from the dataset nor be to sinus like so that all fine structure is lost. Figure 3.7 shows the probability densities for native and false complexes for the three scoring functions PROCOS uses.
With these distributions it is now an easy task to calculate a probability that a complex, not included in the testdata, belongs to either of the two groups represented by different probability densities according to one of the three functions. For a score S it is calculated using Bayes' theorem [33]:

$$p(N \mid S) = \frac{p\,(N) \cdot p(S \mid N)}{p(N) \cdot p(S \mid N) + p(F) \cdot p(S \mid F)} \tag{3.7}$$

Here the probability $p(N \mid S)$ is calculated that a complex belongs to the class of native complexes N according to the score S. The probability densities $p(S \mid N)$ and $p(S \mid F)$ of a given score value S given the native or false distribution can be read out from the corresponding graph in Figure 3.7. The priors $p(N)$ and $p(F)$ are set to 0.5. More details about this choice are given in section 3.3.

3.2.7 Some Ideas for Combining the three Scores to a Single Probability

It is clear that several effects are responsible for the formation of complexes in nature. It has been shown by Murphy et al. [16] that scoring can be improved considerably by combining the information of several scoring functions. Therefore it is advisable to combine different scores into the prediction of PROCOS. This means to combine the scores of the three different scoring functions into one probability output, which will be called PROCOS prob-

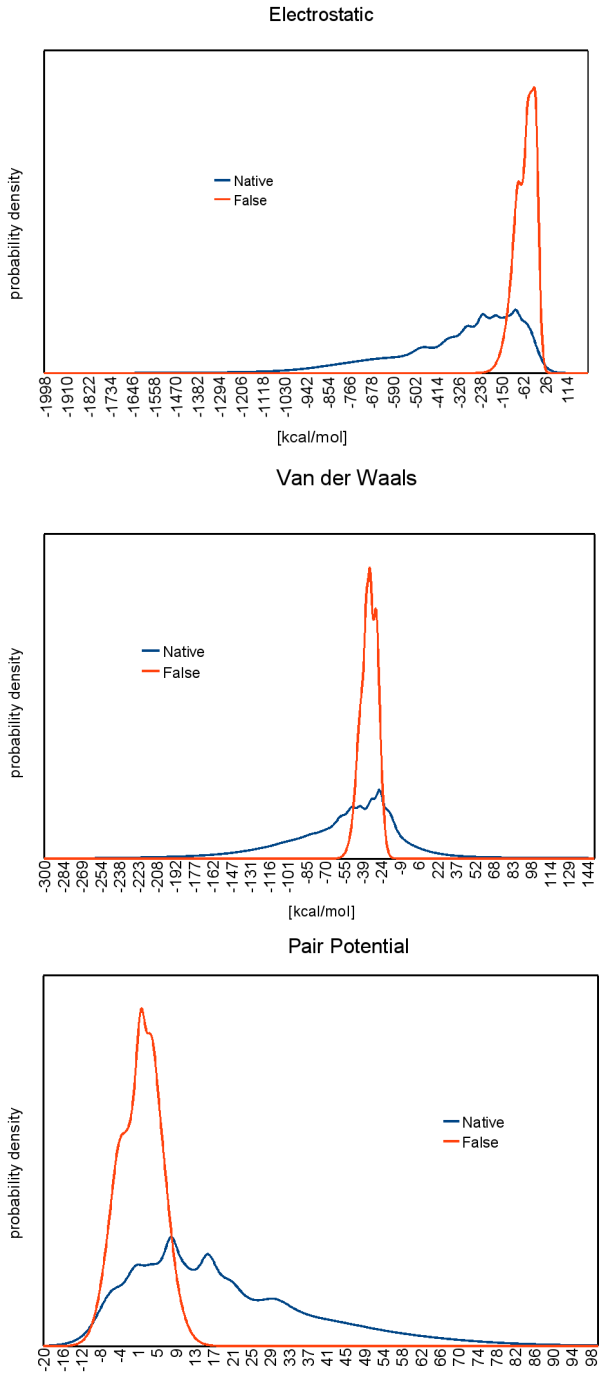


Figure 3.7: Probability densities of the native and false complexes for the three described scoring functions. Electrostatic and van der Waals are energies, and therefore the more negative the value is the merrier the complex, which can be seen from the positions of the native and the false distribution. For the pair potential this is vice versa, as positive values are given to amino acid pairs that are likely to be near each other in the interface. The rescaling explained in section 3.2.5 makes the distributions better comparable.

ability in the sequel. Several ideas to do such a combination were tested in the course of this work. The following subsections will explain them:

Combined Score I (CS1)

The most obvious way to combine the three scores to one probability estimate is to modify equation 3.7 in a way that it can handle several scores:

$$p(N \mid S_{global}) = \frac{p(S_{elec} \mid N) \cdot p(S_{vdw} \mid N) \cdot p(S_{pair} \mid N)}{p(S_{elec} \mid N) \cdot p(S_{vdw} \mid N) \cdot p(S_{pair} \mid N) + p(S_{elec} \mid F) \cdot p(S_{vdw} \mid F) \cdot p(S_{pair} \mid F)} \tag{3.8}$$

Note that the priors were left out in this formula as they do not have an effect when set to 0.5. This formula calculates the probability to belong to the class of native complexes according to all three scoring functions.
Despite this approach seems to be easy and clear, there is a theoretical problem with it. This kind of property combination is only usable for statistically independent scores. It is clear that, for example, the dependency between the electrostatic energy and the pair potential is quite high, as the frequency to find certain amino acids in the interface depends to a considerably degree on their electrostatic interaction. Therefore it was necessary to combine the scores in a way, that their dependencies would not be overrepresented.

Independent Component Analysis (ICA)

One possibility to get dependent signals independent is Independent Component Analysis (ICA). ICA is a statistical tool to solve the Blind Source Separation problem (BSS). A comprehensive introduction is given in the book of Hyvärinen and Oja [34]. ICA is able to recalculate from mixtures of different signals the original independent signals. These independent signals

may be different conversations on a cocktailparty. With microphones in the room only mixtures of all the conversations can be recorded. ICA could in this case calculate the original individual dialogs from the mixtures recorded from the microphones.
In the case of PROCOS, one could interpret the density distributions as mixtures of some underlying unknown properties that are statistically independent. ICA would find the distributions of these properties and it would be possible to use equation 3.8 to calculate probabilities. Figure 3.8 shows the independent probability densities that were calculated by ICA. Since the input signals for ICA have to be 1D vectors, the values from the native and the false distributions were put next to each other in that vector.

However, even though the theoretical idea of this approach seems to be very good, simply looking at the resulting distributions in Figure 3.8 raises doubts about the usability of this method. The problem is that in the resulting plots no more native or false distributions exist but only independent sources of them. The goal was to make the three scoring functions independent, which might be a good idea. But independent sources of native and false signals do not serve the purpose anymore to distinguish between the two classes but between two other unknown classes that nobody is interested in. This happens because ICA is an unsupervised method to find directions in the data that have highest variance. It tries to separate the data in the best way but not necessarily separates native and false datapoints.

To use ICA one step earlier in the process and apply it directly to the scores as they are shown in Figure 3.6 would make it possible to apply ICA separately to the native and false datasets. However, this data is not something that could be called a signal but a list of somehow randomly distributed numbers. As there are no dependencies in these numbers but only in their distributions

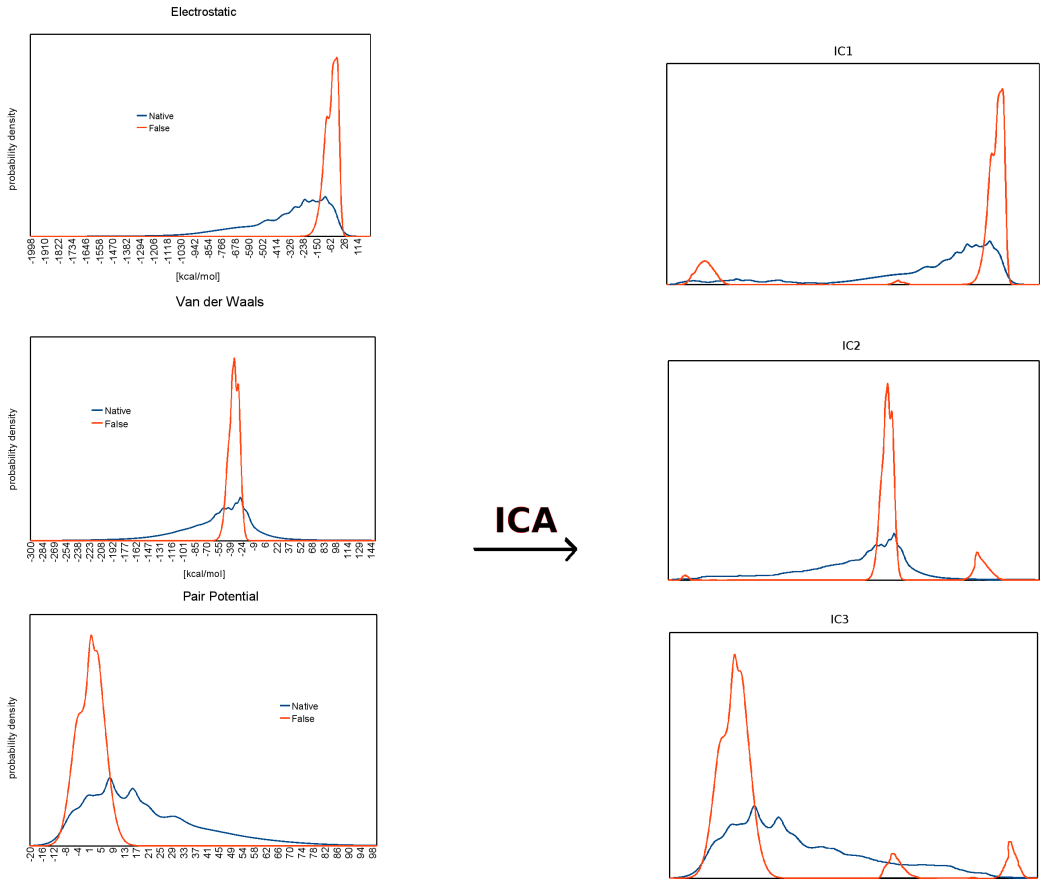


Figure 3.8: Original distributions of the scores to the left and resulting independent components (ICs) of the distributions to the right are shown. The ICs are similar to the original densities and it is visible that some sort of mixing (or rather, in this case, demixing) of the distributions took place. In the new plots no titles, labels and legend can be given, as it is unclear how they would be called after the ICA transformation, which, in the end, leads to no meaningful interpretation of the ICs.

the application of ICA on this data is not meaningful either.
In addition, the presented problem of making three scoring functions independent is actually to low dimensional for typical ICA application.

Combined Score II (CS2)

To avoid ICA and nevertheless eliminate dependencies in the data an approach was developed, which will be called Combined Score II (CS2) here. In this case, the combination of the three scores is not done in the formula but already in the graph. The individual score values of the datasets are plotted in a 3D graph, each dimension representing one scoring function (Figure 3.9). From this plot again a probability density was calculated using equation 3.6 modified for three dimensions. That is a replacement of the scalar x by a vector $v(x, y, z)$ and a new interpretation of the mean μ_n and the variance σ_n, which are calculated from the m nearest neighbors in space in this case. In this case, m was set to 200. Using equation 3.7, the 3D probability density could then be used in the same way as explained before to calculate a probability for a given complex to be native.

Support Vector Machine (SVM)

Despite the CS2 has no statistical problems and provides reasonable results, a second method was developed to deduce a probability to be native from the three measured scores of a given complex: A Support Vector Machine (SVM) was trained with the scores of the datasets of native and false complexes. For the calculation the libSVM library [35] was used. Normally a SVM learns from the given data of two classes a model. This model is then used to classify a new datapoint (the scores of a complex in our case) into one of the classes. However, PROCOS aims for a probability to belong to a class as

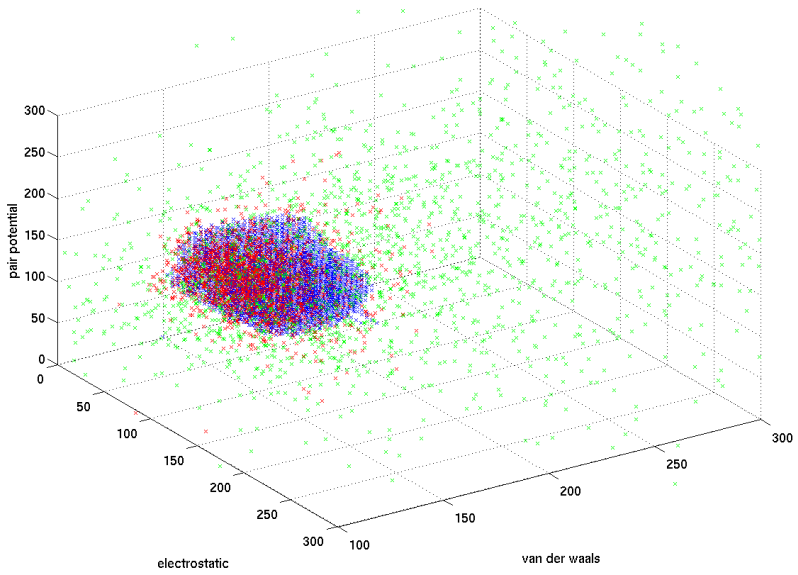


Figure 3.9: 3D plot of Ndat and Fdat1 for all three scoring functions. Native complexes are colored green, false complexes red. The blue surface marks the position where a complex is assigned a probability of 50 % to be native according to equation 3.7. Note that rescaled values are shown at the axes and that only a cut-out of the whole plot is shown to make the small neighborhood of false complexes better visible.

it is very unrealistic to predict the membership of complexes so absolutely. Therefore, the prediction output of the SVM was not used, but the decision values were written into a file and used to produce probability densities in the same way as explained for the scores above. By this it is possible to obtain probabilities as for the CS2 based on a SVM. More details about this approach are given in section 3.3.

3.2.8 Using CAPRI Data as False Distributions

When looking at the distributions of the false complexes in Figure 3.7 it is noticeable how narrow they are compared to the distributions of the native complexes. This effect is probably due to the fact, that the false complexes from Fdat1 were all produced using the same docking program. That means that they were optimized in the same way, which makes them potentially very similar with respect to their energies. For this reason it would be much more realistic to have a false dataset of complexes that comes from different methods. The best resource for such a dataset is probably the CAPRI scoring data (see Appendix). As these complexes were docked by different groups they surely do not have a bias from one special energy optimizing method. Despite the docking was done to obtain near native complexes the vast majority of the structures is not recognized as acceptable from the CAPRI criteria and can be used as false complexes. Since CAPRI data should as well serve as test data for PROCOS only 25% arbitrarily chosen incorrect complexes per target (2194 structures) were used to generate the false probability density. This dataset is called Fdat2 in the following. More details and the resulting curves are presented in the following section.

3.3 A General Overview

This section is going to explain the current version of PROCOS to provide an understandable insight into the used data, the underlying concepts and the interpretation of the results. The next section 3.4 will focus on the program structure of the project.

PROCOS is a webserver that calculates for a given complex a probability like measure to be native. In contrast to scores often used for analyzing complex structures the calculated probabilities offer the advantage of providing a fixed range of expected values. Judgments are based on distributions of properties derived from a large database of native and false complexes. For complex analysis PROCOS uses these property distributions of native and false complexes together with a support vector machine (SVM). In the sequel of this section the datasets will be presented, the used properties (scoring functions) will be explained and the calculation of a probability to be native by using an SVM will be described in detail.

The underlying idea of PROCOS is to classify complexes based on Bayes's theorem [33], which is used to calculate the probability p that a complex with a global score value S belongs to the class of native complexes N:

$$p(N \mid S) = \frac{p\,(N) \cdot p(S \mid N)}{p(N) \cdot p(S \mid N) + p(F) \cdot p(S \mid F)} \tag{3.9}$$

For the calculation estimates of the probability distributions $D_N = p(S|N)$ and $D_F = p(S|F)$ of the property S for the two classes N and F of native and false complexes are required. Although it is possible to formulate *a priori* assumptions on these distributions, the extraction of this information from known complex structures is more robust. Therefore, native complexes were taken from the Mintz database (Ndat), which contains 2541 non homologous native protein complexes [30]. A meaningful antipode of false complexes was

taken from CAPRI scoring data (Fdat2) as detailed below. For each of these complexes the values of three analysis functions were calculated: Intermolecular electrostatic energy (e), intermolecular van der Waals energy (v) and the score of an intermolecular amino acid based pair-potential (k) [31]. The e, v, and k values obtained for each complex in the sets of native and false complexes were used to train a support vector machine (SVM) with two classes. In this case the property S is related to the position of an individual complex relative to the separating hyperplane of the SVM model. Next, using these data probability distributions were obtained for the two classes N and F. Figure 3.10 gives an overview of the procedure which is detailed below.

Finding reasonable values for the *a priori* probabilities $p(N)$ and $p(F)$ that a complex belongs to the class of native complexes N or to the class of false complexes F is a difficult task that depends on several factors such as the docking algorithm used, the system under investigation, etc. As an approximation $p(N) = p(F) = 0.5$ was used. This does, of course, not at all reflect the real proportion between the amount of true solutions and all theoretically possible conformations. However, it would be meaningless to select some other arbitrarily chosen values as long as there are no facts available resulting in more reasonable estimates for the priors. This affects the results in a way that the so called "probabilities" are not real probabilities to be native structures. To obtain somewhat more realistic priors one could scan the solutions of typical docking runs for the fraction of native and non native complexes. For example, the numbers of near native and false complexes of the recent CAPRI scoring competitions could be used for this purpose. This would lead to priors $p(N) = 0.062$ and $p(F) = 0.938$. However, it should be noted that these are no general values and therefore, in this work priors of $p(N) = p(F) = 0.5$ were used.

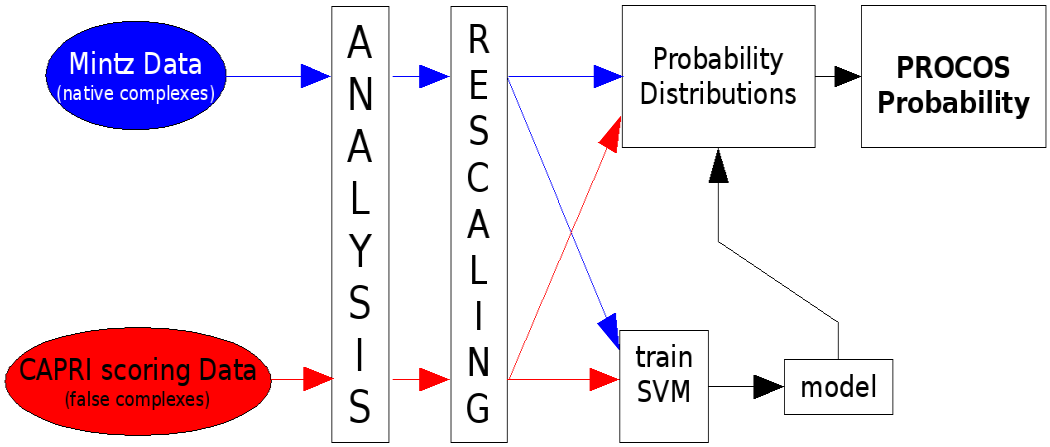


Figure 3.10: Overview of the work-flow to obtain probability distributions for native and false protein complexes: Protein complexes from the Mintz database [30] are used as native complexes. False complexes were taken from erroneous results of the CAPRI scoring competition. For all complexes three different analysis functions were used, namely van der Waals energies, electrostatic energies, and amino acid wise pair potential scores. Resulting values were rescaled for reasons of data comparison. A support vector machine (SVM) was trained with the different scores and a measure related to the distance of every complex to the separating hyperplane was calculated. These data were used to calculate a new set of probability distributions for the two classes N and F. The data flow of native and false complexes is symbolized by blue and red arrows, respectively.

For this approach it is necessary to obtain a reasonable set of false complexes. For creating this set one cannot simply join two proteins in an arbitrary way since the resulting complexes would be extremely unrealistic. For a realistic set, false complexes are needed that do not exist in nature but are, nevertheless, optimized in a way that they could theoretically exist. As a possible solution to this problem already existing decoys from targets of the last CAPRI scoring competitions (T29, T32, T35, T36, T37_1, T38, T39, T40_CA, T41) that were generated by many different predictor groups using a variety of different algorithms were taken (see Appendix). Of those, 25% arbitrarily chosen complexes per target (2194 structures) that were marked as incorrect according to the CAPRI criteria were used for the calculation of the probability distributions of the false complexes (Fdat2). This approach ensures that the resulting distributions are not biased towards a single algorithm used for calculating the structures. The remaining 75% of the data was later used for testing PROCOS. Note, that for targets 37 and 40 two evaluations were performed by CAPRI. For T37 this was done due to high symmetry between the two chains in the ligand of T37 and their close proximity to each other and the interface. For target 40 there are two possible interfaces at opposite sides of the receptor (see CAPRI homepage for details [36]). However, to not overuse the structures of these targets they were used only once for the generation of probability distributions. The so obtained probability distributions for the false complexes represent a meaningful antipode to the group of the native complexes.

Visualization Through Probability Distribution Plots

As the above scoring functions are very diverse in their physical meaning, rescaling of the individual functions was performed for easier visual inspec-

	elec	vdw	pair
$rescaled_0$	0 kcal/mol	0 kcal/mol	0
$rescaled_{1000}$	-1548.24 kcal/mol	-340.18 kcal/mol	114.36

Table 3.3: Conversion factors for the rescaling of the scoring functions. The table shows the original values for the rescaled scores of 0 and 1000. Inserting these values in Equation 3.5 converts an arbitrary value. Note that the pair potential has no unit.

tion. Therefore, for all data the zero point for each function was set to the point where this function adopts a value of zero. By going in the direction of more favorable values a maximum number of 1000 was assigned to the point where the probability density values for the distributions of both the native and false complexes approached a value of zero i.e. they were both below 0.1 % of the largest obtained probability density value of this function. Using the same step size and the same cutoff criteria a rescaling was also performed in the opposite direction. Note that the rescaling is different from that explained in the previous section. In this case, using equation 3.5 the parameters from Table 3.3 have to be used.

From the rescaled data for each analysis function probability distributions were obtained for the groups of native and false complexes according to equation 3.6. The parameter m (number of neighbors that are considered per gaussian) was set to 200.

The resulting rescaled probability distributions are shown in Figure 3.11. Analysis of the diagrams shows that in all cases distinct differences were obtained between the distributions of the native and false complexes. For reasons of comparison also distributions obtained from near native complex structures of the latest CAPRI scoring competitions were included in green.

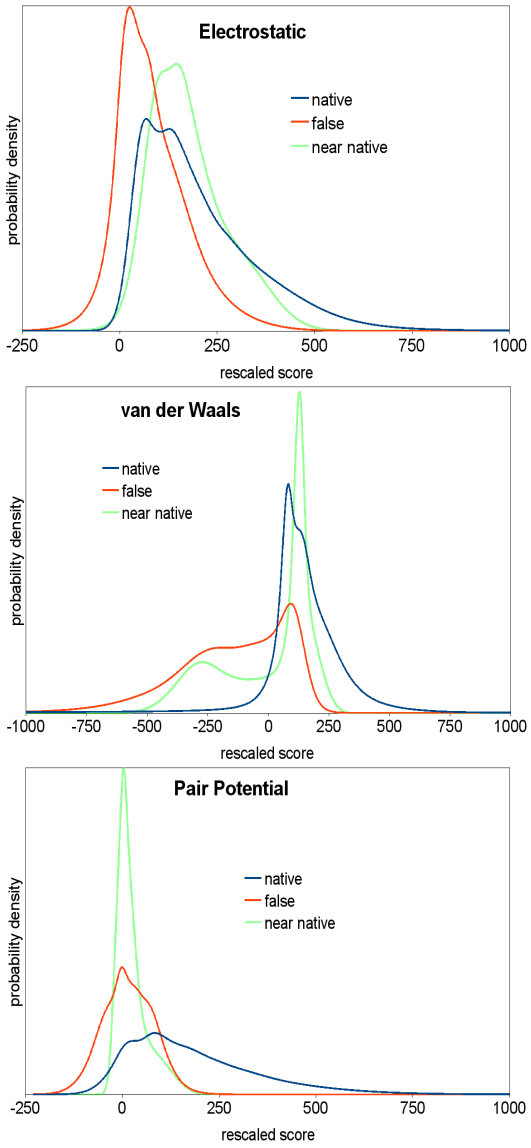


Figure 3.11: Probability distribution plots for electrostatic energy (top left), van der Waals energy (top right) and knowledge based amino acid wise pair potential scores (bottom middle). The curves for the native complexes are plotted in blue, those for the false complexes in red. For reasons of comparison also distributions obtained for the near native structures of the CAPRI test data are included (green). All values are rescaled, see Methods section for details.

Note, that the latter distributions were not used for any calculations.

Calculation of Probabilities with an SVM

To combine the three calculated scores to one global probability measure, an SVM was trained using the libSVM library [35]. For training, the e, v and k values obtained from the complexes of Ndat and Fdat2 were used. In all cases a kernel function with a radial basis was used.
The standard output of a SVM is a yes/no-answer. In our case the SVM decides whether the complex belongs to the group of the native complexes or not. However, as mentioned before, the aim of PROCOS is to calculate a probability like measure that a complex belongs to the class of native complexes. For this, after training, a measure related to the distance of every complex to the separating hyperplane (decision value) is computed. Based on these data probability distributions are calculated as described above. Figure 3.12 shows the corresponding distributions for native and false complexes.
For a newly investigated complex the e, v and k values are calculated and based on these data, the position relative to the separating hyperplane is calculated according to the previously learned model. Using equation 3.7 and the distributions D_N and D_F shown in Figure 3.12 the PROCOS probability measure that this complex belongs to the class of native complexes is computed.

User Interface

To admit all users an easy access to PROCOS, a web interface was implemented (http://compdiag.uni-regensburg.de/procos), which allows the analysis of a binary protein complex to be uploaded as a pdb-file. After parsing the input data, values of the above mentioned analysis functions are calcu-

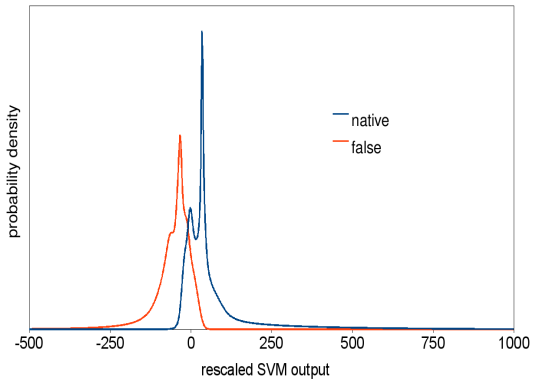


Figure 3.12: Probability distributions of the obtained SVM model. The distributions of native and false complexes are plotted in blue and red, respectively.

lated and displayed together with the corresponding probability distribution plots and the actual values marked by colored bars within it (see Figure 3.13). These data is provided as additional information to the calculated probability measures.

For ranking, of course, it is not useful to only analyze one single complex. Normally, the output of docking algorithms contains hundreds or thousands of complexes that have to be ranked. Therefore, PROCOS is also able to handle up to 150 complexes when they are submitted in one file (maximal file size 32 MB) as different models. In this case no graphical output is given but only a list of all complexes giving the PROCOS probability as well as the single probabilities and the scoring values for each complex. If no chains are selected by the user, the first two chains in the PDB-file are selected automatically. To submit several complexes, their PDB-description has to be in one file, separated by a line with the word "MODEL". The models can be numbered (MODEL 1 ... MODEL 2 ... MODEL 3 ...).

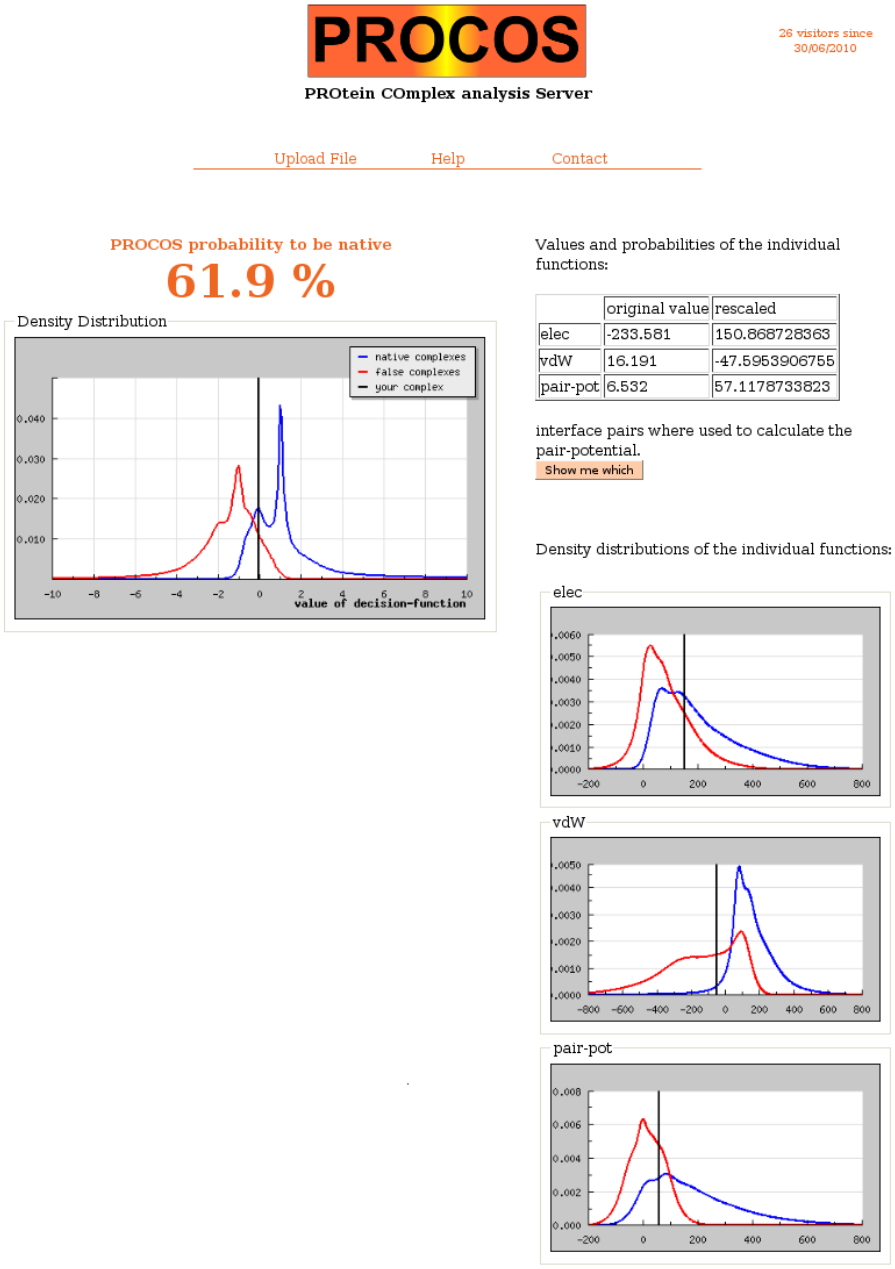


Figure 3.13: Cut-out of the results page of PROCOS when only one complex is analyzed.

3.4 PROCOS in Detail

This section explains the program structure of PROCOS in detail and try to give all necessary information to anybody who will further develop the algorithm. Currently, the whole program is located on the nfs-directory of the Computational Diagnostics group at the University of Regensburg. All used files of PROCOS can be found in /nfs/compdiag/www/htdocs/procos/. This directory is as well mounted on the server and is accessible via http://compdiag.uni-r.de/procos/. The heart of PROCOS is an algorithm called "intermol" that reads in a pdb-file of a protein complex and calculates the three scores explained in the previous section. Around intermol a user interface was written, that handles the file upload, calculates the SVM-probabilities and displays the results to the user.

3.4.1 PHP-Scripts

index.php:
This is the start page which is loaded automatically when typing in the above mentioned address in a web browser. This is achieved by adding the line "DirectoryIndex index.php" into the file *.htaccess*. For this script some other php-files are required that are all explained well in the source code and understandable written (*func.php*, *cleanbuffer.php*, *head.php*, *title.php*, *menu.php*). The html form on this site submits three variables to the file *ergebnis.php*: "datei" (the uploaded pdb-file), "chains" (the selected chains in that file) and "token" (an automatically created variable to avoid multiple uploads of the same file).

ergebnis.php:
Here the uploaded pdb-file is saved on the server with the command

"move_uploaded_file" to Prog/Scoring/PDBs/. Then the script *first_filter.php* removes all lines from the file that do not start with "ATOM" or "MODEL", and counts the number of atoms and models in the file. Then the file is splitted into one file per model and the lines with the "MODEL"-numbering are deleted (this is because the program *reduce.exe* is not able such statements in the code). If there are several models, the script *calc_prob.php* is started, which will be explained later. In case only one model was uploaded, the script *chain_filter.php* is started to remove all chains that are not meant to be analyzed. Then the script *scorerun.php* is started.

scorerun.php:
This is the main analysis file for the case that only one complex (one model) is uploaded. As all steps are explained in the program, only the most important parts will be mentioned here. In the beginning, the file Prog/Scoring/Dat/*verzeichnisse* is changed. This is needed for intermol, to write its outputs to the correct directory. While the calculation runs, an animation of a protein complex is shown. This is done in the included *file index_fake.php*. The meaning of the three different commands "ob_start()", "system($scoring)" and "ob_end_clean()" to start intermol is to prevent text output to the terminal from intermol. When the calculation went well, the outputs are rescaled, probabilities are calculated and rounded and finally the SVM-probability is calculated with the program *svm.php*. In case that at least one of the three scores is outside the range of the training data a SVM-probability can still be calculated, however, it will not be displayed to avoid false interpretations. In the end the *file erg_layout.php* is started to present the output.

calc_prob.php:
As mentioned before, a special program is started when the uploaded file contains several complexes (models). The program is well commented and

in principle similar to *scorerun.php*. The presentation of the results is in this case included in the end of the program and consists only of a list of the score values of the uploaded complexes.

3.4.2 Intermol

Intermol is an algorithm, written in C++ that reads in a pdb-file of a complex and calculates its intermolecular electrostatic energy, van der Waals energy and its pair potential according to Wolowski et al. [31]. The compiled file is called *start* and can be found in Prog/Scoring/. The files with the source code are located in Prog/Scoring/Prog/ as well as the main program *intermol.cpp*. All parameters intermol needs can be found in Prog/Scoring/Dat/.

intermol.cpp:

This is the main script, which calls all necessary routines in sequence. First, all H-atoms are removed and then added again by the program reduce.exe. Intermol needs H-atoms to calculate the correct energies and as labeling for H-atoms is not standardized it is saver to remove eventually existing H-atoms first to be sure that they are labled in the same way. Note that *reduce.exe* does not work properly when the pdb-file contains a MODEL-statement. Therefore, these statements were deleted after splitting the uploaded file. It is possible to disable *reduce.exe* by setting the corresponding variable to 0 in Prog/Scoring/Dat/*parameterzusatz*.

Next, the program *suche_stelleAB* is started. It goes line for line through the pdb-file and reads in the characters in column 21 and 72. These are the positions where the chain identifiers are localized (standard pdb: 21, HADDOCK output: 72). If the positions are both empty (whitespaces) or both contain characters, an errormessage is written to the error file and intermol is aborted. The output of the program are not the chain identifiers

but only the position of them in the file (21 or 72).

The program *zeilen_zaehlen.cpp* counts the number of atoms (lines) for both chains individually and saves them in the pointer "zeilenzahl".

Next, the program *zeilen_lesen.cpp* reads in the atom number, the atom name, the residue name, the residue number and the three coordinates of every atom. As the two chains are handled separately, the whole information is saved in a three dimensional array[chain][linenumber][one of the seven mentioned values]. There are several validations included in this program to test if atom and residue names can be handled from intermol and warning or error messages are written in case the read data failed the tests.

To save computational time, especially for big complexes, the program *calpha_preselect.cpp* is written. All interactions calculated later on in intermol have values significantly different from zero solely for atoms that are located in the neighborhood of the other protein in the complex. Therefore, all residues whose CA-atoms have a distance greater than 20 Å from any CA-atom in the other protein are removed from the array containing the structure information.

Next, the program *parameter_lesen.cpp* reads in the charge parameters for atoms in different amino acids from the file Prog/Scoring/Dat/*parameter*.

After that, the table of scores for the pair-potential (Prog/Scoring/Dat/*score_tabelle*), which was developed from Wolowski et al. [31] is read in from the program *scores_lesen.cpp*.

Finally the actual calculation starts by calling the program *calc.cpp*. After loading several parameters from the file Prog/Scoring/Dat/*parameterzusatz* a dual loop is started, running over all pairings of atoms between the two proteins of the complex. For every pair the distance between the atoms is calculated with the program *calc_dist.cpp*, and then, if the corresponding

distances are small enough, depending on the cutoff parameters of the various scoring functions *calc_Eint.cpp*, *calc_PairPot.cpp* or both of them are started to calculate the scores.

In the end, a fourth score, pair_mean, is calculated which is simply the mean pair potential per interface pair. However, this value was never used for PROCOS analysis. All four scores are then written into files and one file including all scores is created as well. It is called *all*. The files are saved in the directory given in the file Prog/Scoring/Dat/*verzeichnisse* after the "Temp" statement.

As already mentioned, the directory Prog/Scoring/Dat/ contains all data and parameters that intermol needs. The file *parameter* contains the charges for the atoms, *score_tabelle* contains the values for the pair potential, *verzeichnisse* contains the directory where intermol writes it output and the location of the used program *reduce.exe*. The file *parameterzusatz* exists to have an easy tool to change parameters for the use of intermol. Here it is possible to set the values for all cutoff parameters in the equations, limits for the output of warnings, whether or not *reduce.exe* should be used and the radii of the atoms.

There is also a program version of intermol that can be used independently of PROCOS to analyze huge numbers of complexes. It is located on /nfs/compdiag/user/procos_save/Scoring/. It is in principle the same as the above explained program. Its additional features are explained in the appendix.

3.5 Testing PROCOS

In this section, all results that were obtained by testing PROCOS are presented and discussed. The testing was performed with the following sets of protein-protein complexes. A label to refere to these datasets in the sequel is given in brackets. First tests were performed with 96 native complexes from the PDB (NativeTest). Then docked complexes from different algorithms were used including decoys from 13 complexes that were docked with HADDOCK (HaddockTest), the remaining 75% of the CAPRI scoring data that was not used for training (CapriTest) and the 40 dimeric complexes from the Dockground Decoyset [37] (GroundTest). To get an impression how PROCOS is doing compared to other established scoring methods, the just mentioned datasets were scored and reranked by HADDOCK, ZRANK, FireDock and DFIRE as well.

For the 96 native complexes of NativeTest it was made sure that they are on the sequence level at most 25% identical to any complex in Ndat. The complexes were analyzed by PROCOS, ZRANK and DFIRE. When scoring native complexes the desired outcome of the algorithm would be the best possible score. Nearer to native than native is not possible. Results show that PROCOS yielded for 87 of these complexes probability values between 100% and 50% and only for 8 of the native complexes lower values were obtained. The average probability value obtained for all native test structures amounts to 85.2%. When further analyzing the complex showing the lowest probability value of 7.9% it becomes apparent that this complex shows very high van der Waals energies indicating a possible problem with the experimental structure determination. There were a few other complexes with low probability values, in these cases this is mostly due to an unfavorable pair potential score. In addition to the global probability values PROCOS pro-

vides the values of the individual analysis functions to allow for a detailed evaluation of the results. These data clearly shows the advantage of using a probability based analysis scheme since the values obtained for a set of very different complexes are directly comparable with each other. When using more conventional scoring schemes like ZRANK and DFIRE one obtains for the same set of 96 native complex structures a range of scoring values between -814 and -14 (ZRANK) and -234 and 301 (DFIRE). For these values it is absolutely not clear how good they are, although all of them are native complexes. This data shows that the scores obtained depend very much on the type of the investigated complex and do not provide a global measure as it is the goal of PROCOS.

However, the normal use of a scoring algorithm is not the evaluation of native complexes but an analysis of docked complexes to filter out those solutions that are near to the native structure. To test PROCOS' performance on such data, 13 complexes from the Benchmark 3.0 [28] dataset were docked with HADDOCK (1ACC_1SHU, 1C3D_1LY2, 1MZN_1ZGY, 1QG4_1A12, 1RGH_1A19, 1SUR_2TRX, 1TGK_1M9Z, 1UDH_2UGI, 2BME_1YZM, 4PEP_1F32, 1A2P_1A19, 1HDN_1F3G and 1BTP_1LU0). In all cases the unbound structures were taken as starting point for the docking. To achieve a considerable number of near native solutions the interface residues on the larger protein were defined as active whereas all surface amino acids of the smaller protein were defined as passive. This additional information leads to a relatively high amount of near native solutions, sometimes over 50% of the structures, which was meaningful for the first test to reveal the principal potential of PROCOS to rank near native complexes more likely in top positions than obviously wrong solutions. For every target 200 solutions were produced with HADDOCK. Then the complexes were ranked by PROCOS,

ZRANK, HADDOCK, DFIRE and FireDock. According to CAPRI criteria explained in the Appendix it was decided, which solutions are near native. Then it was simply counted how many near native complexes were ranked by each algorithms in the top 10, top 20 and top 50 positions. Table 3.4 shows the resulting numbers. The last line in the table gives the total number of near native solutions for the corresponding target and the total number of solutions (always 200 in this case). Comparing the achieved numbers of near native solutions in the top 10 ranked complexes it is clear that HADDOCK ranking performs best with the highest number of near natives found for 5 of the 13 targets. However, it has to be stated that HADDOCK has an advantage in this case as the complexes are docked with HADDOCK and, therefore, are structurally optimized in a way that the scoring algorithm is built for. Nevertheless, PROCOS ranks for 4 target most near native complexes within the top 10 ranks, which is equally well as FireDock and better than ZRANK and DFIRE. Especially for target 1QG4_1A12, which has the lowest number of near native solutions (12) and therefore is most difficult to score, PROCOS finds 4 of the 12 near native complexes within the top 10 solutions. This is an outstanding result especially when compared to the other algorithms, which partly only find 3 near natives within the top 50

	1ACC_1SHU					**1C3D_1LY2**					**1MZN_1ZGY**				
	PROCOS	ZRANK	HA	DFIRE	FD	PROCOS	ZRANK	HA	DFIRE	FD	PROCOS	ZRANK	HA	DFIRE	FD
top 10	4	5	9	8	6	3	5	0	2	10	10	10	10	10	10
top 20	9	9	17	16	12	5	10	1	5	15	19	20	20	20	20
top 50	25	27	33	42	36	19	30	8	16	35	47	50	50	50	49
near natives	94 of 200					89 of 200					111 of 200				
	1QG4_1A12					**1RGH_1A19**					**1SUR_2TRX**				
	PROCOS	ZRANK	HA	DFIRE	FD	PROCOS	ZRANK	HA	DFIRE	FD	PROCOS	ZRANK	HA	DFIRE	FD
top 10	4	1	0	0	1	1	1	7	4	5	7	7	5	2	1
top 20	4	3	0	0	2	1	5	13	10	11	13	16	7	3	2
top 50	5	5	3	3	4	6	17	33	25	26	28	28	22	12	15
near natives	12 of 200					116 of 200					52 of 200				
	1TGK_1M9Z					**1UDH_2UGI**					**2BME_1YZM**				
	PROCOS	ZRANK	HA	DFIRE	FD	PROCOS	ZRANK	HA	DFIRE	FD	PROCOS	ZRANK	HA	DFIRE	FD
top 10	0	0	3	2	1	4	4	6	8	8	2	6	4	5	2
top 20	3	1	6	3	5	10	10	12	15	13	4	10	7	8	6
top 50	14	12	13	9	11	29	27	33	35	35	13	22	19	20	17
near natives	69 of 200					85 of 200					56 of 200				
	4PEP_1F32					**1A2P_1A19**					**1HDN_1F3G**				
	PROCOS	ZRANK	HA	DFIRE	FD	PROCOS	ZRANK	HA	DFIRE	FD	PROCOS	ZRANK	HA	DFIRE	FD
top 10	3	2	2	1	4	0	0	1	0	2	10	2	5	9	5
top 20	5	5	5	2	8	1	1	2	2	3	15	8	12	15	10
top 50	12	10	12	6	18	4	2	4	5	7	26	14	25	33	23
near natives	29 of 200					14 of 200					62 of 200				

	1BTP_1LU0				
	PROCOS	ZRANK	HA	DFIRE	FD
top 10	6	2	7	6	5
top 20	14	2	14	9	7
top 50	24	10	17	18	18
near natives	55 of 200				

Table 3.4: Analysis of the 13 complexes from the HaddockTest dataset. The Table shows the number of near native solutions found in the top 10, top 20 and top 50 ranked solutions for PROCOS, ZRANK, HADDOCK (HA), DFIRE and FireDock (FD). The last line gives the number of near native structures found in total within the 200 solutions.

ranked complexes. This is a very good result for PROCOS as it obviously keeps up with well established scoring algorithms.

Next, a more realistic test set with regard to the fraction of near native complexes was used: The CapriTest dataset which includes the remaining 75% of the incorrect solutions together with the near native structures of the Capri scoring data that were not used to train PROCOS. As can be seen in Table 3.5 there are much more solutions per target and very view near native solutions. Only 4 out of 1049 for Target 39 for example. That, of course, makes it very hard for the algorithms to rank near native solutions within the top 10 complexes. Nevertheless, the Capri management accepts only 10 complexes to be submitted and the challenge for the scorers is to get some near native structures within these 10. This is actually reasonable, as the goal with the scoring would be to reduce the number of complexes a molecular biologist has to study to understand some biological behavior. Investigating more than 10 complexes is not feasible within a reasonable amount of time. Looking into the table it is obvious that the goal of finding near native complexes within the top 10 ranked solutions often fails, especially when there are only very few native complexes. Therefore, to get a better overview, the numbers of near native solutions found in the top 5%, top 10%, top 20% and top 50% of the total number of structures per target are given as well. Even in this realistic testset PROCOS performs very well. For target 29 and target 41 it has the most near native solutions within the top 10 complexes of all four scoring algorithms. Comparing the other lines of the table it is clear that PROCOS achieves similar results as the other algorithms. Note, that no values were available for target 37 for FireDock, as the FireDock server was not able to analyze this complex.

	T29				**T32**				**T35**			
	PROCOS	ZRANK	DFIRE	FD	PROCOS	ZRANK	DFIRE	FD	PROCOS	ZRANK	DFIRE	FD
top 10	1	0	0	0	0	0	0	0	0	0	0	0
top 5%	58	28	26	44	0	0	0	0	0	0	0	0
top 10%	86	69	66	80	0	0	1	0	0	0	0	0
top 25%	92	99	109	115	0	0	9	0	0	0	2	0
top 50%	108	132	137	135	6	6	15	1	0	0	2	0
near natives	143 of 1607				15 of 386				2 of 351			
	T37_1				**T37_2**				**T39**			
	PROCOS	ZRANK	DFIRE	FD	PROCOS	ZRANK	DFIRE	FD	PROCOS	ZRANK	DFIRE	FD
top 10	0	0	1	-	0	0	0	-	0	0	0	0
top 5%	0	0	3	-	0	0	0	-	0	0	0	0
top 10%	1	5	3	-	0	1	0	-	0	0	0	1
top 25%	9	9	13	-	1	1	5	-	0	0	0	2
top 50%	20	20	25	-	5	2	6	-	0	0	0	3
near natives	34 of 843				7 of 843				4 of 1049			
	T40_CA				**T40_CB**				**T41**			
	PROCOS	ZRANK	DFIRE	FD	PROCOS	ZRANK	DFIRE	FD	PROCOS	ZRANK	DFIRE	FD
top 10	2	7	0	8	2	2	5	0	5	1	4	1
top 5%	4	31	17	50	47	27	17	16	22	8	14	15
top 10%	12	73	34	74	57	40	25	33	40	36	19	44
top 25%	85	181	99	221	64	60	28	72	101	128	43	122
top 50%	201	246	148	328	73	67	36	81	192	186	178	235
near natives	346 of 1568				88 of 1568				295 of 893			

Table 3.5: Analysis of the complexes from the CapriTest dataset. The Table shows the number of near native solutions found in the top 10 ranked solutions for PROCOS, ZRANK, DFIRE and FireDock (FD) in the first line. Since every target has a different total number of complexes the other rows give the number of near native solutions found in the same percentage of complexes (not absolute numbers) to simplify comparison between different targets.

A third test with a greater number of targets was performed with the 40 dimeric complexes from the Dockground Decoyset [37] (GroundTest). This dataset not only comprises more complexes from different targets, but also no targets from it were used for training PROCOS. This is a more independent test as it could be argued that the complexes used in the CapriTest may introduce some bias in the testing procedure. This can not be the case for GroundTest.

Groundtest actually contains more than 40 targets, but these contain more than two proteins per complex. Since PROCOS was only designed for dimeric protein complexes these targets were left out.

Due to a greater amount of analyzed data PROCOS was only compared to ZRANK and DFIRE in this case. The results of the obtained rankings are given in Tables 3.6 and 3.7, showing the number of near native solutions found in the top 10, top 20, and top 50 structures. For method comparison the number of near native structures within the 10 top ranked structures was counted. In case that the same number of structures was obtained the following line of the tables (top 20 ranked solutions) was evaluated. In case that no distinction was achieved by the top 2 lines the compared methods were considered as equal. Inspection according to these rules shows that in comparison with ZRANK in 23 cases PROCOS performs better, in three cases both methods perform equal and in 14 cases ZRANK outperforms PROCOS. In comparison with DFIRE, PROCOS performs in 29 cases better, in 4 cases equal and in 7 cases worse. This is a good result for PROCOS, especially when taking into account that PROCOS was developed from the ground during three years and only uses three different scoring functions to archive this performance.

	1avw_A_B			1bui_A_C			1bui_B_C			1bvn_P_T		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	4	2	0	5	1	0	0	1	0	7	0	10
top 20	5	3	0	8	4	0	5	4	0	8	2	12
top 50	8	5	5	10	9	0	6	6	0	10	6	12
near natives	10 of 110			10 of 110			10 of 110			12 of 110		
	1cho_E_I			1dfj_E_I			1e96_A_B			1ewy_A_C		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	8	5	8	6	0	1	0	1	1	5	3	0
top 20	9	7	12	7	0	3	0	2	6	6	7	0
top 50	13	10	15	7	0	7	2	7	10	8	9	4
near natives	15 of 110			10 of 109			10 of 110			10 of 110		
	1f6m_A_C			1fm9_A_D			1g6v_A_K			1gpq_A_D		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	0	0	0	9	2	8	1	2	0	6	3	0
top 20	0	1	0	12	4	13	2	3	0	6	5	1
top 50	2	3	0	12	7	13	6	8	0	8	6	8
near natives	10 of 110			13 of 110			8 of 108			10 of 110		
	1gpw_A_B			1he1_A_C			1he8_A_B			1ku6_A_B		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	0	1	9	3	3	1	1	1	0	4	3	0
top 20	6	2	14	3	5	1	1	1	0	8	4	0
top 50	9	3	17	6	9	8	1	1	0	8	7	0
near natives	18 of 110			13 of 110			1 of 101			10 of 110		
	1ma9_A_B			1nbf_A_D			1oph_A_B			1ppf_E_I		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	8	3	6	7	6	0	2	4	0	3	6	0
top 20	9	5	8	7	8	0	5	5	0	6	8	1
top 50	10	7	9	8	10	4	7	9	3	10	10	8
near natives	10 of 110			10 of 110			10 of 110			10 of 110		

Table 3.6: *explanation see Table 3.7 next page*

	1r0r_E_I			**1s6v_A_B**			**1t6g_A_C**			**1tmq_A_B**		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	5	2	4	2	2	0	9	4	9	4	0	7
top 20	6	5	8	2	4	0	19	8	19	4	1	10
top 50	8	9	12	2	4	2	47	25	48	6	4	10
near natives	13 of 110			4 of 104			66 of 110			10 of 110		
	1tx6_A_I			**1u7f_A_B**			**1ugh_E_I**			**1w1i_A_F**		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	0	0	0	4	6	0	4	0	10	3	2	0
top 20	1	2	0	7	8	0	5	0	12	3	3	0
top 50	4	3	0	9	9	0	7	1	12	4	4	0
near natives	10 of 110			11 of 110			12 of 110			4 of 104		
	1wq1_R_G			**1x3d_A_B**			**1yvb_A_I**			**2a5t_A_B**		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	2	2	0	6	2	8	5	6	4	0	0	0
top 20	3	2	0	9	2	10	6	10	6	0	0	0
top 50	5	7	4	9	6	10	10	11	10	0	0	1
near natives	12 of 110			10 of 110			11 of 110			1 of 101		
	2bkr_A_B			**2btf_A_P**			**2ckh_A_B**			**2fi4_E_I**		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	2	0	0	6	3	0	3	5	2	3	2	0
top 20	5	1	1	8	4	1	4	7	7	4	2	0
top 50	7	2	10	9	9	8	6	9	9	11	6	2
near natives	11 of 110			10 of 110			10 of 110			11 of 110		
	2goo_A_C			**2sni_E_I**			**3fap_A_B**			**3sic_E_I**		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
top 10	0	0	0	4	2	0	4	4	0	4	6	0
top 20	0	0	0	6	2	3	6	4	0	6	7	1
top 50	4	4	7	8	4	6	7	6	0	9	10	4
near natives	10 of 110			10 of 110			10 of 110			10 of 110		

Table 3.7: Reranking results of 40 additional targets from GroundTest. For every target the number of near native structures in the top 10, top 20 and top 50 ranked solutions is shown. Note that the number of near native solutions found is not directly comparable to the numbers from the CapriTest reranking as the percentage of near native structures is considerably higher in the GroundTest decoyset.

To obtain these results it was not necessary to divide the complexes into different groups e.g. enzyme-inhibitor, antibody-antigen and others as proposed in the literature [27, 38] and adapted by several scoring approaches (e.g. [17, 18]). This enhances the usability and general applicability of PROCOS.

Another possibility to compare the performance of different scoring algorithms is to calculate the Receiver Operating Characteristic (ROC). This is to plot the percentage of false positive hits against the true positive hits. In the case of scoring: Which fraction of the near native complexes is recognized correctly for a given percentage of incorrect solutions that are wrongly classified as near native? Figure 3.14 shows ROC-curves for the CapriTest and the GroundTest datasets. For GroundTest PROCOS outperforms the two other methods significantly. Except for a very strict cut-off where only some few results are considered to be positive (bottom left corner of the graph) and the other extreme situation where nearly all structures are defined as positive (top right corner of the graph), PROCOS finds a lot more true positive structures for a certain amount of false positive structures then the other two methods. For the CapriTest the curve for FireDock is shown in an extra figure (Figure 3.15), because this algorithm was not able to score the targets 37_1 and 37_2. Therefore, the curve is not directly comparable to the other curves as it is based on a lower number of complex structures. Nevertheless, it performs very well, and even if a direct comparison is not possible it is obvious that FireDock analyzes the Capri-structures best of all tested algorithms. However, it has to be discussed how useful ROC-curves are in general for comparing scoring algorithms. The final goal of the ranking is to find all or at least many of the near native solutions in top positions. It does not help for further analysis of the solutions if most of the good com-

plexes are higher ranked than most of the incorrect ones. This is actually the reason why CAPRI accepts only a submission of 10 complexes. It is not feasible to analyze more than these 10 complexes for a biologist who needs the correct structure for his studies. Therefore, the only really interesting part of a ROC-curve in this comparison is the bottom left-hand corner of the plot. In this corner one can see how many of the top ranked structures really are near native. Looking into the plots it appears that all algorithms except DFIRE for the CapriTest seem to perform quite similar according to the curves. Therefore, it is probably better to compare scoring algorithms by means of the before presented ranking lists.

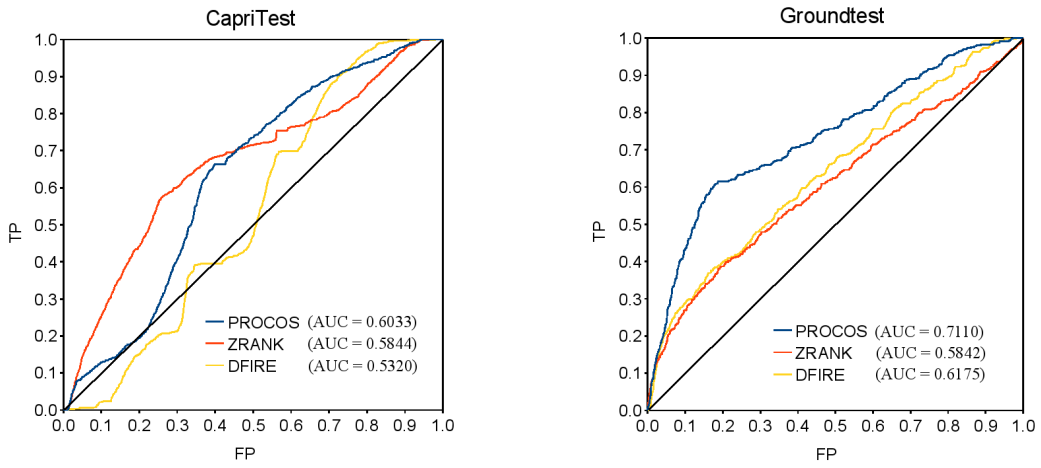


Figure 3.14: ROC-curves for PROCOS, ZRANK and DFIRE on the two datasets CapriTest and Groundtest. Especially for CapriTest all algorithms show very poor results. However, in Groundtest PROCOS clearly outperforms the two other methods.

When comparing the ROC-curves and the ranking lists it is possible to get the impression of contradictorily results. This is especially distinct for DFIRE in CapriTest. The ROC-curve in the beginning of the graph is very bad. That means that the best ranked 10% of the complexes do hardly contain any true

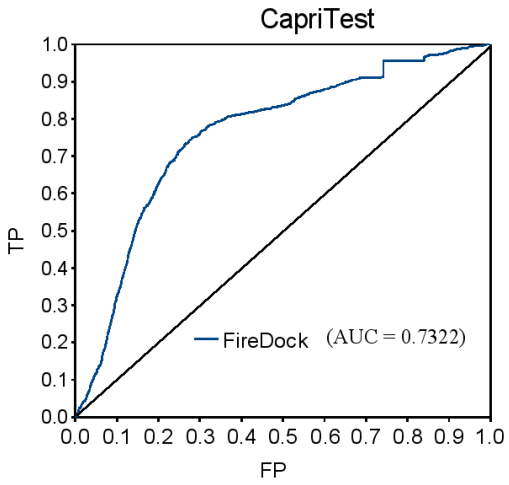


Figure 3.15: ROC-curve for FireDock on the CapriTest dataset. This curve is not included in Figure 3.14 as FireDock was not able to analyze targets 37_1 and 37_2. Therefore, the database is different and the curves not directly comparable.

positive results. However, in the ranking list (Table 3.5), there are some targets where DFIRE perform very well even in the top 10 complexes (especially T40_CB and T41). The explanation for this behavior is the following: For the ROC-curves, the scores of all solutions from the different targets are analyzed together in one list. In this example, DFIRE obtains scores of below -1000 for 472 complexes. This are according to DFIRE the best complexes of the whole CapriTest. However, these complexes belong only to two targets, T35 and T39, and contain only two near native solutions. This is the reason that the ROC-curve for DFIRE looks so bad in the bottom left corner of the graph.

The data that is produced from scoring algorithms for so many protein complexes is difficult to handle and even more difficult to interpret. Therefore, a third way to present the results and to compare them to other algorithms was implemented to focus on the data from a slightly different side. For

this presentation the average scoring results from the near native solutions and the average results from the 25% worst ranked solutions were compared. The values are shown in Tables 3.8 and 3.9. For all targets, the average scores (ZRANK, DFIRE) and probability values (PROCOS) of the near native solutions according to CAPRI criteria were calculated. The corresponding values are shown in the first two/three columns. The data contained in the two/three columns to the right were calculated by taking the mean of the 25% worst solutions of a target according to the measures calculated by PROCOS, ZRANK and DFIRE. Values for the CAPRI targets 36 and 38 are not shown as they do not contain near native complexes.

Analysis of the CapriTest targets in Table 3.8 shows that for the near native structures of targets 32, 35 and 39 the average probability values amount to very small numbers of 2.7%, 11.6% and 0.5%, respectively. For these targets also the total number of near native structures is relatively small with numbers of 15, 2 and 4. This indicates that these are quite difficult targets and that the obtained near native solutions are still not optimal. When comparing the probability values of the near native structures with those of the 25% worst solutions a clear gap is visible that allows setting of a global threshold to safely remove a considerable subset of the wrong structures. The corresponding average score values obtained by ZRANK for the near native solutions of the various targets are between -144.1 and 1162.7 and for the 25% worst solutions a range of 42.6 to 1749.7 is obtained. It is reasonable to argue that these values are more difficult to interpret than the PROCOS probability values. One of the advantages of the PROCOS values is that they are by definition within well defined limits between 0% and 100%. Although, as explained before, due to the issue of defining appropriate priors, a PROCOS probability value cannot yet be interpreted as a real probability

that a given complex structure is close to its native form. Currently PROCOS uses the approximation of $p(N) = p(F) = 0.5$. Using the already mentioned possibility to take the CAPRI structures to calculate the priors would not change the analysis procedure in principle but shift the obtained probabilities for being in the class N of near native complexes to lower values. Such probability measures would of course be nearer to real probabilities that native complexes are found, but would depend significantly on the arbitrarily chosen dataset they were derived from.

Looking at the Haddocktest targets in Table 3.8 it can be seen that for the near native structures average probability values between 75.0% nd 43.4% were obtained. These are considerably higher values than obtained for most of the CAPRI targets showing the increased average quality of the decoys due to the inclusion of interface information in the docking routine. Also the range of probability values is relatively small indicating that most of the near native structures of the various targets are of comparable quality. For the worst 25% of all solutions probability values from 55.6% to 6.2% were found. This reflects the fact that by the inclusion of additional interface information the generation of totally wrong solutions is mostly prevented. This behavior can be nicely followed using PROCOS. ZRANK calculates for the near native structures scores between -75.2 and -337.0, whereas for the worst 25% a range between -51.3 and -195.33 is obtained. This also demonstrates that PROCOS provides a first step towards a global measure for complex analysis that should allow the comparison of complexes from different targets with each other.

Finally, when analyzing the average PROCOS probability values for the near native structures of Groundtest (Table 3.9), a range between 55.9% and 0.3% is obtained while the corresponding range for the worst 25% solutions

complex	**average results near native solutions**		**average results 25% worst solutions**	
Capritest	PROCOS	ZRANK	PROCOS	ZRANK
29	0.6017	-104.11	0.0020	42.61
32	0.0274	73.2	0.0024	912.06
35	0.1163	1162.66	0.0025	1358.02
37_1	0.3269	-49.89	0.0026	610.25
37_2	0.2353	37.92	0.0026	610.25
39	0.0051	364.37	0.0023	1749.67
40_CA	0.2777	-30.77	0.0030	1509.34
40_CB	0.6957	-144.11	0.0030	1509.34
41	0.1884	20.89	0.0031	1679.57
Haddocktest	PROCOS	ZRANK	PROCOS	ZRANK
1ACC_1SHU	0.4858	-88.05	0.2652	-55.50
1C3D_1LY2	0.4488	-75.15	0.2654	-51.31
1MZN_1ZGY	0.5794	-336.99	0.1448	-153.67
1QG4_1A12	0.4910	-296.92	0.0623	-195.33
1RGH_1A19	0.4341	-166.81	0.2280	-122.80
1SUR_2TRX	0.7498	-179.30	0.5559	-126.15
1TGK_1M9Z	0.6376	-159.18	0.5205	-109.49
1UDH_2UGI	0.6126	-144.45	0.3728	-91.14
2BME_1YZM	0.7039	-198.13	0.5435	-117.60
4PEP_1F32	0.6266	-167.72	0.4239	-107.09
1A2P_1A19	0.5749	-147.21	0.3286	-99.46
1HDN_1F3G	0.6784	-164.75	0.4249	-114.36
1BTP_1LU0	0.6605	-196.79	0.4202	-137.93

Table 3.8: For all targets, the average scores (ZRANK) and probability values (PROCOS) of the near native solutions according to CAPRI criteria were calculated. The corresponding values are shown in the first two columns. The data contained in the two columns to the right were calculated by taking the mean of the 25% worst solutions of a target according to the measures calculated by PROCOS and ZRANK. Values for the CAPRI targets 36 and 38 are not shown as they do not contain near native complexes.

is between 0.1% and 0.2%. These data show that the probability values obtained for these two sets of structures do not overlap and setting of a global threshold to safely remove a considerable subset of the wrong solutions seems feasible. The advantage of such a global threshold is that it may be selected *a priori* independent of the investigated target. The corresponding average score values obtained by ZRANK for the near native solutions of the various targets are between 37.7 and 3253.8 and for the 25% worst solutions a range of 1047.5 to 28.64 is obtained. For DFIRE a range between -30.7 and 204.8 is computed for the near native solutions while for the 25% worst solutions a range between -18.9 and 213.7 is calculated (Table I.). For both ZRANK and DFIRE considerable overlap exists between the score values obtained for the near native structures and the 25% worst structures, which makes the setting of a target independent threshold for selection purposes more difficult.

In conclusion of this analysis one can say that the classification of complexes with a probability like measure as done by PROCOS has the following advantages:
(i) A probability value is in principle more meaningful for deciding whether a complex is native or not compared to a score where it is unclear which threshold should be used to decide whether a complex should be selected for further analysis. However, it is clear that in the current implementation PROCOS is only able to calculate probability-like measures.
(ii) A considerable sub-section of the false complexes can be eliminated from further analysis by setting of an appropriate threshold *a priori*.
(iii) It is possible to compare the results from different targets with each other.
(iv) PROCOS also performed well in the task of reranking existing decoys of docking runs as shown on the Dockground decoy set.

complex	average results near native solutions			average results 25% worst solutions		
	PROCOS	ZRANK	DFIRE	PROCOS	ZRANK	DFIRE
1avw_A_B	0.2062	623.10	-20.27	0.0021	1522.81	-14.00
1bui_A_C	0.2085	455.49	-10.27	0.0017	1805.63	-10.02
1bui_B_C	0.0107	623.70	-16.36	0.0018	1814.42	-12.01
1bvn_P_T	0.2440	1050.17	-30.70	0.0020	1707.82	-16.88
1cho_E_I	0.2981	283.83	-20.73	0.0020	1047.46	-12.60
1dfj_E_I	0.3636	3253.77	-11.83	0.0010	2786.44	-8.64
1e96_A_B	0.0030	627.08	-18.29	0.0019	1597.88	-9.51
1ewy_A_C	0.0617	356.14	-15.42	0.0017	1801.21	-11.13
1f6m_A_C	0.0028	970.85	-12.21	0.0020	1878.92	-11.41
1fm9_A_D	0.1830	903.57	-28.02	0.0017	1841.94	-9.92
1g6v_A_K	0.0100	399.81	-10.85	0.0018	1566.24	-10.41
1gpq_A_D	0.1101	431.02	-14.61	0.0022	1220.22	-9.22
1gpw_A_B	0.0124	1672.01	-20.09	0.0018	1944.96	-8.77
1he1_A_C	0.0248	799.12	-14.87	0.0015	1791.32	-10.07
1he8_A_B	0.5593	118.37	-10.28	0.0011	2489.49	-11.42
1ku6_A_B	0.0230	733.64	-13.19	0.0019	1564.13	-12.77
1ma9_A_B	0.0871	1269.32	-28.27	0.0010	2864.80	-15.72
1nbf_A_D	0.1435	220.88	-10.09	0.0014	1913.58	-4.78
1oph_A_B	0.0340	510.61	-14.54	0.0014	2016.66	-11.83
1ppf_E_I	0.2221	37.68	-18.52	0.0024	1277.76	-11.47
1r0r_E_I	0.0185	476.91	-17.25	0.0023	1304.52	-10.23
1s6v_A_B	0.0229	261.11	-8.82	0.0015	1704.70	-4.80
1t6g_A_C	0.0786	1142.53	-28.15	0.0020	1843.32	-18.91
1tmq_A_B	0.0508	990.57	-27.09	0.0020	1621.02	-15.85
1tx6_A_I	0.0111	1223.19	-18.24	0.0018	2158.75	-18.00
1u7f_A_B	0.1799	366.85	-17.46	0.0020	1729.67	-15.33
1ugh_E_I	0.0353	1388.10	-26.31	0.0020	1636.86	-8.07
1w1i_A_F	0.0106	539.27	-8.38	0.0010	2546.49	-10.50
1wq1_R_G	0.0030	1151.19	-14.41	0.0016	2426.31	-10.40
1xd3_A_B	0.1712	625.74	-17.48	0.0019	1685.97	-6.01
1yvb_A_I	0.0811	65.21	-19.16	0.0021	1630.27	-11.49
2a5t_A_B	0.0026	1583.81	204.80	0.0012	2394.10	213.73
2bkr_A_B	0.0077	1217.61	-14.90	0.0019	1638.06	-6.19
2btf_A_P	0.1889	584.26	-18.13	0.0013	1799.41	-12.04
2ckh_A_B	0.0694	160.95	-12.39	0.0018	1528.02	-4.72
2fi4_E_I	0.1164	649.15	-14.86	0.0022	1144.50	-12.47
2goo_A_C	0.0027	714.86	-22.68	0.0021	1205.66	-15.85
2sni_E_I	0.0835	627.15	-15.50	0.0022	1317.28	-9.56
3fap_A_B	0.0953	201.55	-11.76	0.0023	1087.03	-10.39
3sic_E_I	0.1224	155.81	-17.21	0.0019	1381.82	-13.95

Table 3.9: For all targets of Groundtest, the average scores (ZRANK and DFIRE) and probability values (PROCOS) of the near native solutions according to CAPRI criteria were calculated. The corresponding values are shown in the first three columns. The data contained in the three columns to the right were calculated by taking the mean of the 25% worst solutions of a target according to the measures calculated by PROCOS, ZRANK and DFIRE.

PROCOS is freely available as an easy to use web server. Processing a pdb-file containing a protein complex it calculates a probability-like measure that this structure belongs to the class of native complex structure. To support the user's decision, the computed values are visualized in a plot which represents the probability distributions of the training data. In future developments we expect further improvements by adding additional analysis functions. Due to the modular concept of PROCOS, this can easily be achieved.

Chapter 4

Docking Applications

4.1 Model of the Saratin-Collagen Complex

Figure 4.1: *Hirudo medicinalis*

A typical example for a docking experiment was performed during this work in collaboration with the Institute of Biophysics and Physical Biochemistry. In this context, the protein structure of the leech protein Saratin was analyzed [39]. As the interaction of Saratin and Collagen is of special interest for a possible protein based drug against coagulation the complex structure of these two proteins was modeled as part of the publication [39]. The first part of this section will give a short insight into the biological background of Saratin and its known behavior in connection with Collagen, whereas the second part will concentrate on the modeling of the complex structure.

4.1.1 Background

Collagen is the main protein of connective tissue. When exposed to blood upon injury, thrombocytes recognize exposed Collagen and start the process of blood coagulation, which prevents the organism from bleeding to death. However, in some cases, for example to prevent heart attacks, it is desired to suppress this mechanism. Saratin, which was isolated from the saliva of the leech *Hirudo medicinalis* (Figure 4.1), could be a powerful therapeutic component to locally prevent coagulation. It is known that Saratin binds to Collagen but the exact mode of interaction has still to be revealed. To identify the binding site of Saratin, NMR spectroscopy was used to search for amino acids with significant chemical shifts for increasing amounts of Collagen in the probes. Those amino acids that, in addition to the chemical shift, are at least 20% solvent accessible were then defined as direct interaction partners. These were the residues Thr8, Glu39, Tyr40, Tyr42, Glu61 and Tyr78. In addition, a 3D structure of Collagen is also required for the construction of a reliable complex model. For the collagen peptide used in this study, $(Gly-Lys-Hyp-(Gly-Pro-Hyp)_{10}-Gly-Lys)_3$, no X-ray or NMR structure was available. As a consequence, a triple-helical Collagen model structure was obtained, employing the specifically dedicated modeling program THe BuScr [40]. From the highly repetitive primary, secondary, and tertiary structure of Collagen, it can be assumed that Collagen exhibits multiple binding sites for Saratin.

4.1.2 Docking

A 3D model of the complex was calculated using the data-driven docking algorithm HADDOCK to gain further insight into the Saratin-Collagen com-

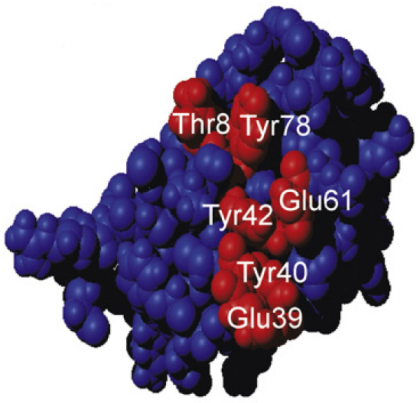


Figure 4.2: Space-filling model of Saratin where the residues participating in binding (according to NMR spectral changes and water accessibility as analyzed by Gronwald et al. [39]) are marked in red. Note that the flexible C-terminus from Thr80 is not shown.

plex formation. For this purpose, the structure of uncomplexed Saratin, together with the information of the binding site from chemical shifts was used. Figure 4.2 shows a space-filling model of Saratin where the residues participating in binding are marked in red. These residues were defined as active residues in the docking run. Because of the unassured binding behavior of Collagen all residues in this protein were defined as passive, allowing all residues of Collagen to potentially participate in binding to Saratin.

In the first step of the docking, 1000 complex structures were calculated by randomization of orientations and rigid-body energy minimizations. Of these, the best 200 in terms of HADDOCK scores were further optimized by simulated annealing and water refinement. The obtained 200 complex structures were sorted according to their HADDOCK scores. All top-scoring complex structures consistently show a very similar binding of Saratin to the N-terminal region of Collagen, indicating that this region is the preferred Saratin binding area. Figure 4.3 shows the best model of the Saratin-Collagen

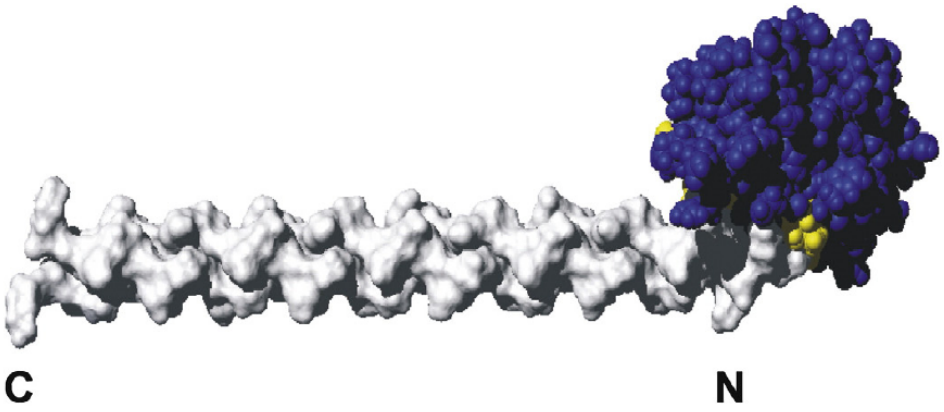


Figure 4.3: Model of the Saratin-Collagen complex in terms of both interaction energy and HADDOCK score. Saratin is bound to the N-terminal part of the triple-helical Collagen model peptide. Saratin is shown in blue, Collagen in light grey. The active residues in Saratin are colored in yellow (hardly visible in the Figure).

complex in terms of both HADDOCK score and intermolecular interaction energy in a space-filling representation. Residues of Saratin that were defined as active residues are shown in yellow. This result is very reasonable especially when taking electrostatic considerations into account. The positively charged side chain of Lys2 of Collagen at the N-terminus is in the model in direct contact with the negatively charged binding site of Saratin mainly represented by Glu61 and Glu62 (see Figure 4.4a). In addition, the binding site of Saratin is surrounded by positively charged residues not directly involved in the interaction as can be seen in Figure 4.4b. However, these amino acids may lead the Collagen strand in the right direction and may further stabilize the bound Collagen strand to the interaction site.

It is important to be aware of the fact, that the obtained complex is not the native structure of the Saratin-Collagen interaction. The model is, with a certain probability, similar to the complex existing in nature and can be used

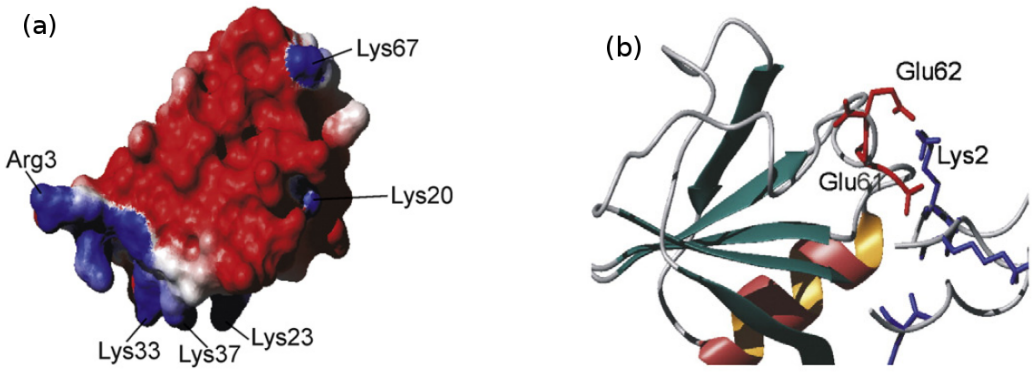


Figure 4.4: (a) Electrostatic surface potentials of Saratin (red: negative, blue: positive) that were calculated with the program MOLMOL [41]. The protein has the same orientation as in Figure 4.2. (b) Detail of the interaction site in a ribbon representation. The negatively charged Lys2 side chain of Collagen shown in blue is sandwiched between the two positively charged glutamate side chains of Saratin shown in red. Of the Collagen strand only a short part of the N-terminus is visible.

as a working hypothesis for further investigations. As mentioned above, no scoring method is able to reliably rank the best solutions in top positions today. However, when several top ranked structures are very similar to each other, as it was the case in this study, this is a strong argument that a near native solution is found. It would be very unlikely that multiple randomized starting structures are optimized to the same energy minimum if it was not the global minimum.

4.2 MIA

Another typical example for the use of docking methods was performed in a collaboration with the Institute of Pathology at the University of Regensburg in a study about the functional inhibition of the MIA protein [42]. Again,

the section is divided into two parts. First, the general background of the studied proteins and the work of the collaboration partners is outlined and then the focus is set on the docking, which is a part of the present work.

4.2.1 Background

The melanoma inhibitory activity (MIA) protein is secreted by melanoma cells. Melanoma is the most aggressive form of skin cancer (see Figure 4.5). Normally, not the skin cancer itself leads to death but the main problem is the early occurrence of metastases. MIA strongly supports the formation of metastases and the study demonstrated that dimerization of MIA is required to start its functional activity. To investigate further on the structure of the MIA dimer the interface regions were predicted with the PreBI modeling software [43]. According to this model the interface is located at the amino acids Tyr31, Arg56 and Arg58 on the one hand and Ser64, Tyr70, Asp72, Leu73 and Ala74 on the other hand. This implies a head to tail linkage of the complex and admits the assumption that even oligomers of higher order may be formed from MIA, which was not further investigated on in this study.

Figure 4.5: *Melanoma malignum*

As it became clear that MIA is probably only active in form of a dimer the question arose if one could find an appropriate inhibitor protein or peptide that prevents the formation of dimers and therefore the tendency of the melanoma to form metastases. This would be a big step in the direction of a new strategy in melanoma therapy. Therefore, peptides that are

generally known to bind to MIA were screened for their potential to prevent MIA dimerization and to induce dissociation of already existing protein dimers. The screening resulted in the dodecapeptide AR71 (sequence: Ac-FHWRYPLPLPGQ-NH_2) to be an appropriate inhibitor for the dimerization process: Western blot analysis showed a strong reduction of the dimer band for pre-incubation of MIA with AR71 in comparison to other peptides. Multidimensional NMR spectroscopy was then used to predict those residues that are probably involved into the binding process. Amino acids with significant chemical shifts that were at the same time at least 20% solvent accessible were Cys17, Ser18, Tyr47, Gly66, Asp67, Leu76, Trp102, Asp103 and Cys106 of MIA.

4.2.2 Docking

HADDOCK came in use to model the 3D structure of the MIA dimer complex. As docking partners two MIA proteins (pdbID: 1HJD) in their unbound state were taken from the Protein Database (PDB) [3]. To lower the quantity of possible solutions the interface predicted from PreBI was given to HADDOCK as AIRs. Residues Tyr31, Arg56 and Arg58 on the first docking partner and Ser64, Tyr70, Asp72, Leu73 and Ala74 on the second docking partner were defined as active residues, all surface residues neighboring the active residues were defined as passive. Analysis of the docking results and especially the violations of AIRs showed that HADDOCK had some problems to find an energetically favorable conformation of the dimer with the given surface information. All 200 resulting structures are shown in Figure 4.6 in a ribbon representation, where only the backbone is visible as a line. The complexes are superimposed according to the Cα atoms of the second docking partner, which is shown in blue. The active residues Tyr70, Asp72,

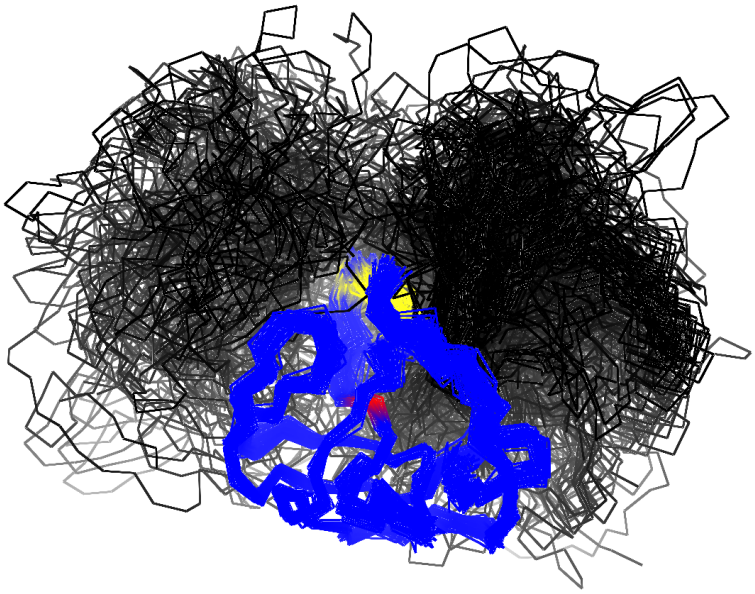


Figure 4.6: Ribbon representation of the first 200 models of the MIA-dimer. The models are superimposed according to the Cα atoms of the second docking partner, shown in blue.

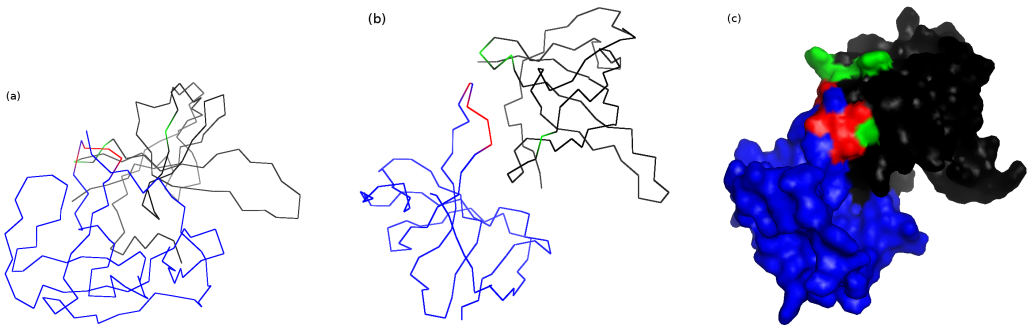


Figure 4.7: Final model of the MIA dimer. First docking partner shown in black and active residues in green, second docking partner shown in blue and active residues in red. (a) Complex from same perspective as in Figure 4.6. (b) Complex from more clear perspective. (c) Complex in surface representation.

Leu73 and Ala74 are colored in cyan, the active residue Ser64 in red. It is notable that Ser64 is lying quite hided inside the protein (yet still solvent accessible), which makes it difficult to access for the docking partner. This results in very high AIR violations for this amino acid. In addition, the resulting structures show a separation into two main clusters (black clouds of backbones). One can assume that each cluster tries to reach Ser64 from a different side and still does not fulfill the AIRs properly. As a consequence, the assumption was reasonable that the prediction of PreBi including Ser64 into the interface was not correct. For this reason, a new docking run was started and Ser64 was defined as passive residue as well. In this case the resulting structures chowed much lower values for the AIR violations, even for the remaining active residues. From the biggest cluster of solutions the complex with the best HADDOCK score (which includes the AIR violations beside different energy terms) was selected as a good model for the MIA dimer. Figure 4.7 shows the chosen solution (a) in the same perspective as Figure 4.6, (b) from a more clear perspective and (c) in a surface representation.

Next, the focus of interest was set on the complex structure of MIA with the inhibitor peptide AR71. Knowledge about the structure would give an additional hint on if it is geometrically supported that the binding of AR71 disturbs the formation of the MIA dimer. As the sequence of AR71 contains an acetyl- and an NH_2-end which are not parameterized in the HADDOCK program it was necessary to define these terminal residues manually. Therefore, the files “parallhdg5.3.pro”, “topallhdg5.3.pro” and “topallhdg5.3.pep” were adjusted and modified accordingly. The CTER parameters were changed to add a NH_2-end and a new block ACET was written for the acetyl-end. All additional necessary atoms, impropers and dihedrals were added together with a label that manual changes have been performed. After this prelimi-

active residue	Cys17	Ser18	Tyr47	Gly66	Asp67	Leu76	Trp102	Asp103	Cys106
run2	114	104	119	64	93	59	169	146	114
run3	110	109	118	78	102	64	x	158	120
run4	98	77	85	39	60	74	172	162	x
run5	106	71	71	44	46	68	x	154	x

Table 4.1: Number of AIR violations for different active residues. An "x" in the table means that this amino acid is not defined as active in this run.

nary work it was possible to dock the peptide to MIA. Here, the supplemental information from the chemical shifts mentioned before was used to define active residues. On MIA no passive residues were defined and all residues of AR71 were defined as passive, as nothing was known about its binding site. Several docking runs were performed to obtain the most probable conformation of the complex. In this case, violations of AIRs were analyzed to identify those active amino acids that are most difficult to fulfill for the docking algorithms. There is a high probability that this kind of residues are showing chemical shifts during complex formation not due to proximity to the peptide but for other reasons. Therefore, the model probably becomes better when leaving out these amino acids. Table 4.1 shows in how many cases of the 200 docking solutions the defined active residues were violated in different docking runs. In run2 all residues with chemical shifts were defined as active. In run3 Trp102 was left out as it had the highest number of violated AIRs (169). In run4 Cys106 was left out. For this residue the violation number was not extraordinary high, but the assumption was reasonable, that this residue draws the peptide away from the other active residues and augments their violation values. Finally, in run5, both residues were left out and it could be shown that not only the AIR violations were lower for most residues, but also was the clustering of the 200 solutions more homogeneous. That means, that the solutions were lying nearer to each other and were indicating one single conformation of the complex which is probably near the native com-

plex structure. In conclusion it can be stated that the AR71 binding site is located in the cleft next to the distal loop of MIA. Geometrical considerations lead to the understanding that AR71 is able to inhibit formation of dimeric MIA molecules and therefore prevent the formation of metastases. Further clinical investigations on mouse models actually showed a reduced formation of metastases on application of AR71 [42].

Chapter 5

Summary and Outlook

In this work dimeric protein complexes were analyzed. The main part of the time was dedicated to the development of a scoring algorithm for docked protein-protein complexes (PROCOS) [25]. The analysis of the docking results was based on score distributions of databases of native and false complexes. This approach is totally new and opens in addition the possibility to study the differences between native structures, near native docking solutions and incorrect docking solutions as shown in Figure 3.11. PROCOS was tested on different datasets and compared to well established scoring algorithms.

Classification of complexes with a probability like measure as done by PROCOS has the following advantages:

(i) A probability value is in principle more meaningful for deciding whether a complex is native or not compared to a score where it is unclear which threshold should be used to decide whether a complex should be selected for further analysis. However, it is clear that in the current implementation PROCOS is only able to calculate probability-like measures.

(ii) A considerable sub-section of the false complexes can be eliminated from

further analysis by setting of an appropriate threshold *a priori*.
(iii) It is possible to compare the results from different targets with each other.
(iv) PROCOS also performed well in the task of reranking existing decoys of docking runs as shown on the Dockground decoy set.

PROCOS is freely available as an easy to use web server. Processing a pdb-file containing a protein complex it calculates a probability-like measure that this structure belongs to the class of native complex structure. To support the user's decision, the computed values are visualized in a plot which represents the probability distributions of the training data. In future developments further improvements are expected by adding additional analysis functions. Due to the modular concept of PROCOS, this can easily be achieved. Furthermore, the underlying datasets of native and incorrect complex structures can be complemented by more structures in the future to give PROCOS an even broader basis.

Another part of this work was to use the docking program HADDOCK to find models for different complex structures that were needed for a better understanding of the behavior of these complexes in their biological environment. In this context the Saratin-Collagen complex was modeled, which is an important first step in the development of a powerful therapeutic component to locally prevent coagulation [39]. Moreover, models of MIA-dimers and MIA with its inhibitor AR71 were obtained, which gave more insight into the processes taking place in the metastasation of skin cancer. These studies may be a first step in the development of a novel skin cancer therapy [42].

During the research on protein interactions and the development of PROCOS

in the scope of this work it is natural that several other, connected questions and ideas arise. It is not possible to follow all of them to their end. However, one prominent idea that was present from the beginning of this work but was never investigated on in more detail shall be discussed here in the end: As mentioned in the first chapter, protein-protein interactions play a major role in cellular processes and both experimental and bioinformatic high-throughput methods like yeast-2-hybrid assays are widely used for obtaining interaction maps. However, since these methods are not always applicable and often contain a considerable number of false positives [44], there is a need for computational approaches to verify or falsify protein-protein interactions that were predicted by other methods. It is important to notice that in this case not the structure of the complex is the focus of interest but only the question if interaction is taking place, no matter how. Since protein-protein interactions are critically dependent on the three-dimensional structures of the individual molecules it seems logical to use this information for judging putative protein-protein interactions. Aloy and Russell [5], for example, have suggested a method to model putative interactions on known 3D complexes to investigate the compatibility of a proposed interaction with this complex. In another approach comparative docking together with the analysis of steric clashes is used to analyze putative interactions [45]. A similar goal could be achieved with PROCOS as well. In the very first thoughts for PROCOS (section 3.2.1) it was tested if one can see a difference in the interaction energies between docked protein complexes of proteins that do interact in nature and protein pairs that were forced to interact by the docking algorithm. Figure 3.3 showed exactly this behavior and Table 3.1 confirmed this trend on a larger database. These first results were also published in [46] and [29]. The goal for an algorithm that predicts protein-protein interaction would not be

to give a probability to a certain complex conformation to be native or to rank several complexes according to their probabilities as PROCOS does. For interaction prediction it would be necessary to calculate the PROCOS probabilities of a huge number of docking results of some protein A with another protein B, take the mean probability of all results and compare it to the mean probability that was achieved from A docked to a third protein C. If such an algorithms works well, one could read from this comparison which of the two proteins B and C is more likely to interact with A. Or, even better, if they do at all interact with A. Investigations like this were not done with PROCOS, but it would be an interesting study to use PROCOS in this way, too.

Appendix A

CAPRI

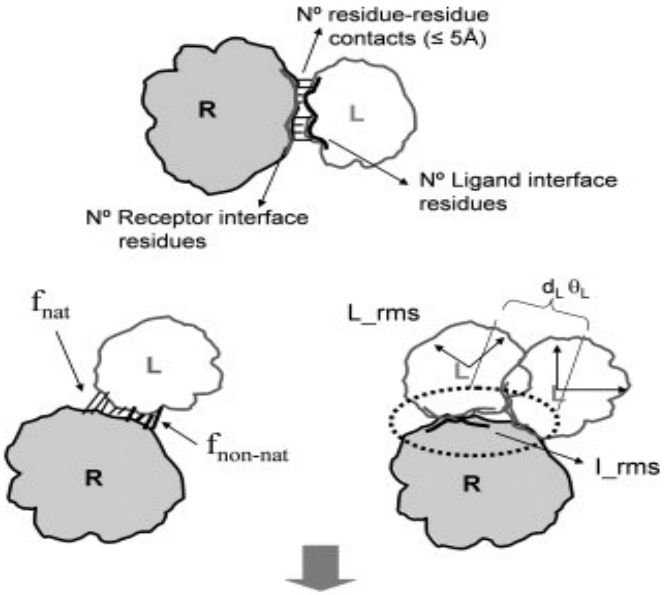
Critical Assessment of PRediction of Interactions is a communitywide experiment on the comparative evaluation of protein-protein docking for structure prediction. Summarizations of the recent CAPRI rounds can be found at Janin et al. [21], Lensink et al. [22] or on the CAPRI homepage [36]. As in the recent years a number of methods have been developed to predict protein-protein interactions CAPRI has been designed as a community wide experiment to assess the progress of those methods. The central question in the event is the following: "If we know the 3D structure of two components of a complex and build a model of their assembly, how reliable and accurate is that model likely to be?"

CAPRI is a blind prediction experiment. Its targets are unpublished crystal or NMR structures of complexes, communicated on a confidential basis by their authors to the CAPRI management. Participant predictor groups are given the atomic coordinates of two proteins that make biologically relevant interactions. They model the target complex with the help of the coordinates and other publicly available data (sequence, mutations etc.), and submit sets of ten models for assessment on the CAPRI web site. In addition, the predic-

tors are invited to upload larger sets that are communicated to scorer groups who evaluate and rank them, and make a separate ten-model submission. After the prediction round is completed, the CAPRI assessors compare the submissions to the experimental structure.

Since CAPRI began in 2001, the experiment has had 22 rounds with one or a few targets per round. Up to now 46 targets have been processed.

For PROCOS the CAPRI scoring data from 9 targets was used: T29, T32, T35, T36, T37_1, T37_2, T38, T39, T40_CA, T40_CB and T41. In the evaluation CAPRI defines four different complex qualities: incorrect, acceptable, medium and high. Table A.1 shows the conditions that are used to classify the docked complexes. In the present work, only a discrimination between false complexes (= incorrect) and near native complexes (=acceptable, medium and high) is done. That means, that a complex is classified as near native if f_{nat} is at least 0.1 and at the same time the L_rmsd is below 10 Å or the I_rmsd is below 4 Å.



Ranking	**Conditions based on CAPRI computed parameters**
High	$f_{nat} \geq 0.5$ AND ($L\_rmsd \leq 1.0$ OR $I\_rmsd \leq 1.0$)
Medium	($f_{nat} \geq 0.3$ AND $f_{nat} < 0.5$) AND ($L\_rmsd \leq 5.0$ OR $I\_rmsd \leq 2.0$) OR $f_{nat} \geq 0.5$ AND $L\_rmsd > 1.0$ AND $I\_rmsd > 1.0$
Acceptable	($f_{nat} \geq 0.1$ AND $f_{nat} < 0.3$) AND ($L\_rmsd \leq 10.0$ OR $I\_rmsd \leq 4.0$) OR $f_{nat} \geq 0.3$ AND $L\_rmsd > 5.0$ AND $I\_rmsd > 2.0$
Incorrect	$f_{nat} < 0.1$ OR ($L\_rmsd > 10.0$ AND $I\_rmsd > 4.0$)

Figure A.1: Schematic illustration of the quality measures used in CAPRI to evaluate predicted models. The following quantities are computed for each target: (1) all the residue-residue contacts between the Receptor (R) and the Ligand (L), and (2) the residues contributing to the interface of each of the components of the complex. For each predicted model the following quantities were computed: the fractions f_{nat} of native and $f_{non-nat}$ of non native contacts in the predicted interface; the root mean square displacement (rmsd) of the backbone atoms of the ligand (L_{rms}), the miss-orientation angle θ_L and the residual displacement d_L of the ligand center of mass after the receptor in the model and experimental structures were optimally superimposed. In addition I_{rms} is computed, the rmsd of the backbone atoms of all interface residues after they have been optimally superimposed. Here the interface residues were defined less stringently on the basis of residue-residue contacts. All tests in this work only distinguished between near native (acceptable, medium, high) and false (incorrect) complexes.

Appendix B

Technical Remarks

B.1 Intermol Without PROCOS

In /nfs/compdiag/user/procos_save/Scoring/ the same program structure for intermol can be found as described in section 3.4. Here the program is not incorporated in a webserver and is easier to use directly from the terminal. Output is by default written to /nfs/compdiag/user/procos_save/Scoring/Temp. To have the possibility to calculate several complexes with one program call the python scrip vieleEint.py is written. The program must be called with one parameter, which is the directory name for the intermol output that will be located in /nfs/compdiag/user/fif01930/EOut/. Then intermol is started sequential for every pdb-file that is written in the file filelocation. Normally, filelocation is a symbolic link to one of the files in the directory Filelocation/.

This program calculates only the score values. To get the probability values of the SVM, other programs have to be used. First the data must be rescaled, which is done in /nfs/compdiag/user/procos_save/Datamanipulation/ and then the SVM can be used, which is located in

/nfs/compdiag/user/procos_save/SVM/. In both directories the shellscript "go" starts the calculations automatically.

B.2 Creating Distribution Plots

To visualize the distributions of the scores of native and false complexes, probability density plots are shown. The datapoints for these plots are calculated in /nfs/compdiag/user/procos_save/Dist/. The python script make_dist.py uses the "all.normiert"-files from /nfs/compdiag/user/procos_save/Datamanipulation/ to calculate the distributions. That should be the two files with the scores of the native and false complexes that will be plotted. The parameters "nachbarn_nuss" and "nachbarn_false" are the number of neighbors of one datapoint (parameter m that are taken to calculate the mean and the variance for the gaussian (see section 3.3). The resulting files (dist_0, dist_1 and dist_2) correspond to the distributions of electrostatics, van der Waals and pair-potential. These files can then be exported into a spreadsheet, for example OpenOffice.org Calc, to draw the plots.

Own Publications

(1) **Fink F**, Merkl R, Gronwald W: Verification of Protein-Protein Interactions by Use of Docking Techniques. In *Proceedings of the NIC Workshop 2007; Jülich.* Edited by Hansmann UHE, Meinke JH, Mohanty S, Zimmermann O, John von Neumann Institute for Computing, Jülich 2007:125-127.

(2) **Fink F**, Ederer S, Gronwald W: Protein-Protein Interaction Prediction. In *Proceedings of the NIC Workshop 2008; Jülich.* Edited by Hansmann UHE, Meinke JH, Mohanty S, Nadler W, Zimmermann O, John von Neumann Institute for Computing, Jülich 2008:209-212.

(3) Gronwald W, Bomke J, Maurer T, Domogalla B, Huber F, Schumann F, Kremer W, **Fink F**, Rysiok T, Frech M and Kalbitzer HR: Structure of the Leech Protein Saratin and Characterization of Its Binding to Collagen. *J.Mol.Biol.* 2008, 381:913-927.

(4) **Fink F**, Ederer S, Gronwald W: Protein-Protein Interaction Analysis by Docking. *Algorithms* 2009, 2:429-436.

(5) **Fink F**, Hochrein J, Wolowski V, Merkl R, Gronwald W: PROCOS: Computational Analysis of Protein-Protein Complexes. *J. Comp. Chemistry* 2011, *in press*.

(6) Schmidt J, Riechers A, Stoll R, Amann T, **Fink F**, Hellerbrand C, Gronwald W, König B, Bosserhoff AK: Dissociation of functionall active MIA dimers by dodecapeptide AR71 inhibits metastasis of malignant melanoma, *submitted*.

Bibliography

[1] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, PHillips DC: **A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis**. *Nature* 1958, **181**:662–666.

[2] Branden C, Tooze J: *Introduction to Protein Structure*. New York: Garland 1991.

[3] **The PDB archieve** [[http://www.pdb.org]].

[4] Zhang Y: **Progress and challenges in protein structure prediction**. *Curr Opin Struct Biol* 2008, **18**:342–348.

[5] Aloy P, Russel RB: **Ten Thousand Interactions for the Molecular Biologist**. *Nat. Biotech.* 2004, **22**:1317–1321.

[6] Young KH: **Yeast Two-Hybrid: So Many Interactions, (in) So Little Time ...** *Biol. Reprod.* 1998, **58**:302–311.

[7] Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla

S, Bickelhaupt E, Lazovatsky Y, DaSilva A, and2 C A Stanyon and2 R L Finley Jr JZ, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM: **A Protein Interaction Map of Drosophila melanogaster**. *Science* 2003, **302**:1727–1736.

[8] Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Séraphin B: **The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification**. *Methods* 2001, **24**:218–229.

[9] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *Proc. Natl. Acad. Sci. USA* 1999, **96**:4285–4288.

[10] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA: **Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques**. *Proc. Natl. Acad. Sci. USA* 1992, **89**:2195–2199.

[11] Gabb HA, Jackson RM, Sternberg MJ: **Modelling protein docking using shape complementarity, electrostatics and biochemical information**. *Journal of Molecular Biology* 1997, **272**:106–120.

[12] Meyer M, Wilson P, Schomburg D: **Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking**. *J Mol Biol 1996* 1996, **264**:199–210.

[13] Moont G, Gabb HA, Sternberg MJ: **Use of pair potentials across protein interfaces in screening predicted docked complexes**. *Proteins* 1999, **35**:364–373.

[14] Fernández-Recio J, Totrov M, Skorodumov C, Abagyan R: **Improving CAPRI predictions: Optimized desolvation for rigid-body docking**. *Proteins* 2005, **58**:134–143.

[15] Kozakov D, Clodfelter KH, Vajda S, Camacho CJ: **Optimal clustering for detectiing near-native conformations in protein docking**. *J Mol Biol* 2005, **89**:199–210.

[16] Murphy J, Gatchell DW, Prasad JC, Vajda S: **Combination of scoring functions improves discrimination in protein-protein docking**. *Proteins* 2003, **53**:840–854.

[17] Li CH, Ma XH, Shen LZ, Chang S, Chen WZ, Wang CX: **Complex-type-dependent scoring functions in protein-protein docking**. *Biophys. Chem.* 2007, **129**:1–10.

[18] Martin O, Schomburg D: **Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines**. *Proteins* 2008, **70**:1367–1378.

[19] Dominguez C, Boelens R, Bonvin AMJJ: **HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information**. *J. Am. Chem. Soc.* 2003, **125**:1731–1737.

[20] Ritchie DW: **Recent Progress and Future Directions in Protein-Protein Docking**. *Curr. Prot. Pep. Sci* 2008, **9**:1–15.

[21] Janin J, Wodak S: **The Third CAPRI Assessment Meeting Toronto, Canada, April 20-21, 2007**. *Structure* 2007, **15**:755–759.

[22] Lensink MF, Méndez R, Wodak S: **Docking and scoring protein complexes: CAPRI 3rd Edition**. *Proteins* 2007, **69**:704–718.

[23] deVries SJ, vanDijk ADJ, Krzeminski M, vanDijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AMJJ: **HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets**. *Proteins* 2007, **69**:726–733.

[24] Jorgensen WL, Chandrasekhar J, Madura JD, Impley RW, Klein ML: **Comparison of simple potential functions for simulating liquid water**. *J. Chem. Phys.* 1992, **79**:926–935.

[25] Fink F, Hochrein J, Wolowski V, Merkl R, Gronwald W: **PROCOS: Computational Analysis of Protein-Protein Complexes**. *J. Comp. Chemistry* 2011, **in press**.

[26] Fink F, Ederer S, Gronwald W: **Protein-Protein Interaction Prediction**. In *Proceedings of the NIC Workshop 2008; Jülich*. Edited by Hansmann UHE, Meinke JH, Mohanty S, Nadler W, Zimmermann O, John von Neumann Institute for Computing, Jülich 2008:209–212.

[27] 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.

[28] Hwang H, Pierce B, Mintseris J, Janin J, Weng Z: **Protein-Protein Docking Benchmark version 3.0**. *Proteins* 2008, **73(3)**:705–709.

[29] Fink F, Ederer S, Gronwald W: **Protein-Protein Interaction Analysis by Docking**. *Algorithms* 2009, **2**:429–436.

[30] Mintz S, Shuman-Peleg A, Wolfson HJ, Nussinov R: **Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions**. *Proteins* 2005, **61**:6–20.

[31] Wolowski VR: **Computational analysis of protein-protein complexes related to knowledgebased predictions of interaction**. *PhD thesis*, Department of Computer Science, University of Hagen, German 2008.

[32] Mie G: **Zur kinetischen Theorie der einatomigen Krper**. *Annalen der Physik* 1903, **11**:657–697.

[33] Cornfield J: **The Bayesian Outlook and its Application,**. *Biometrics* 1969, **25**:617–642.

[34] Hyvärinen A, Oja E: *Independent Component Analysis: Algorithms and Applications*. Helsinki: John Wiley and Sons 2000.

[35] Chang CC, Lin CJ: **LIBSVM : a library for support vector machines**. *Software available at http://www.csie.ntu.edu.tw/c̃jlin/libsvm* 2001.

[36] **CAPRI homepage** [[http://www.ebi.ac.uk/msd-srv/capri/]].

[37] Liu S, Gao Y, Vakser IA: **Dockground protein-protein docking decoy set**. *Bioinformatics* 2008, **24(22)**:2634–2635.

[38] Chen R, Mintseris J, Janin J, Weng Z: **A protein-protein docking benchmark**. *Proteins* 2003, **52**:88–91.

[39] Gronwald W, Bomke J, Maurer T, Domogalla B, Huber F, Schumann F, Kremer W, Fink F, Rysiok T, Frech M, Kalbitzer HR: **Structure of the Leech Protein Saratin and Characterization of Its Binding to Collagen**. *J.Mol.Biol.* 2008, **381**:913–927.

[40] Rainey JK, Goh MC: **An interactive triple-helical collagen builder**. *Bioinformatics* 2004, **20**:2458–2459.

[41] Koradi R, Billeter M, Wtherich K: **MOLMOL: a program for display and analysis of macromolecular structures**. *J. Mol. Graphics* 1996, **14**:51–55.

[42] Schmidt J, Riechers A, Stoll R, Amann T, Fink F, Hellerbrand C, Gronwald C, Knig C, Bosserhoff AK: **Dissociation of functionally active MIA dimers by dodecapeptide AR71 inhibits metastasis of malignant melanoma** 2010, **submitted**.

[43] **PreBI modeling software** [[http://pre-s.protein.osaka-u.ac.jp/prebi/]].

[44] von Mering C, Krause R, Snel B, Cornell M, Oliver S, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**:399–403.

[45] Cockell SJ, Oliva B, Jackson RM: **Structure-based evaluation of *in silico* prediction of protein-protein interactions using comparative docking**. *Bioinformatics* 2007, **23**:573–581.

[46] Fink F, Merkl R, Gronwald W: **Verification of Protein-Protein Interactions by Use of Docking Techniques**. In *Proceedings of the NIC Workshop 2007; Jülich*. Edited by Hansmann UHE, Meinke JH,

Mohanty S, Zimmermann O, John von Neumann Institute for Computing, Jülich 2007:125–127.

[47] DeLano WL: **The PyMOL Molecular Graphics System**. *DeLano Scientifi, San Carlos, CA, USA* 2002, :http://www.pymol.org.

List of Figures

1.1 http://www2.chemistry.msu.edu/Portraits/images/Berzelius3c.jpg 9

1.2 http://www.aloeveraibs.com/wp-content/uploads/2008/08/ aminoacid-struc.jpg 11

1.3 http://www.vectorsite.net/tpchem_07_03.png 11

1.4 http://academic.brooklyn.cuny.edu/biology/bio4fv/page/prot_struct-4143.JPG 12

1.5 http://www.fermentation-biotec.de/downloads/aprotinin.gif 16

1.6 http://riboworld.com/50s/50sbau.jpg 16

1.7 http://earth.usc.edu/ zheqians/images/final_project_csci653/3D_ energy_landscape.bmp 18

2.1 Figure created with GIMP 2.4.7 22

3.1 http://compdiag.uni-r.de/procos/ 26

3.2 Figure taken from [26] 27

3.3 Figure taken from [46] 29

3.4 Taken from [31] 33

3.5 Figures created with OpenOffice Calc 36

3.6 Figures created with OpenOffice Calc 38

3.7 Figures created with OpenOffice Calc 40

3.8 Figures created with OpenOffice Calc 43

3.9 Figure created with Matlab 45

3.10 Figure created with OpenOffice Draw 49
3.11 Figures created with OpenOffice Calc 52
3.12 Figures created with OpenOffice Calc 54
3.13 http://compdiag.uni-r.de/procos/ 55
3.14 Figure created with OpenOffice Calc 71
3.15 Figure created with OpenOffice Calc 72

4.1 http://www.abc.net.au/reslib/200704/r138958_475481.jpg 79
4.2 Figure taken from [39] 81
4.3 Figure taken from [39] 82
4.4 Figures taken from [39] 83
4.5 http://www.upol.cz/uploads/RTEmagicC_MALIGNES_ MELANOM. jpg.jpg 84
4.6 Figure created with Pymol [47] 86
4.7 Figures created with Pymol [47] 86

A.1 Figure taken from [22] 97

Acknowledgment

Many people deserve gratitude for standing on my side during these four years and supporting me with scientific discussions, a good working atmosphere and personal help:

- Prof. Dr. Wolfram Gronwald for a very friendly and competent care. You had a lot of time and understanding for me. Thanks for a very good scientific experience.

- Prof. Dr. Rainer Spang for financial support and the constructive critics on my work.

- Prof. Dr. Elmar Lang for some giving conversations and a lot of help with the universitary administration.

- My office mate Matthias Klein. It was always comfortable to work with you and your Swedish flair made me feel a bit at home. It really was a good time.

- Stefanie Kohl for her open happiness and female atmosphere you brought into our office. I appreciated it a lot to have you there the last year.

- "My" interns Stefan Ederer and Jochen Hochrein for a lot of help in programming and analysis of data.

- All present and former members of the Computational Diagnostics Group especially for the good lunch talks: Benedict Anchang, Inka Appel, Dr. Stefan Bentink, Dr. Maria Alice Bertolim, Sabine Botzler, Peter Butzhammer, Dr. Julia Engelmann, Tully Ernst, Daniela Herold, Dr. Christian Hundsrucker, Dr. Juby Jacob, Phillip Knollmüller, Christian Kohler, Dr. Claudio Lottaz, Matthias Maneck, Katharina Meyer, Giuseppina Moffa, Mohammad Javad Sadeh, Marian Thieme.

- The members of the Institute of functional Genomics for the nice trips and the Christmas parties.

- Thanks to my parents who made it possible for me to study and encouraged me in my way.

- Many thanks to my wife Malin. You had to bear a lot during these years and it was not easy but we got through it. I need you more than I realize.

- Finally I am sure that even my children Emil and Alva contributed to a better success of this work. Their laughter often made me happy again when my brain was stuck in unsolvable science.

Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe des Literaturzitats gekennzeichnet. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe eines Promotionsberaters oder anderer Personen in Anspruch genommen. Niemand hat von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.
Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prfungsbehörde vorgelegt.

Regensburg, den 06.04.2011

.....

Florian Fink