Computational Analysis of Protein-Protein Interactions

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von
Diplom-Biologe
Sam Ansari

Saarbrücken 2007

Tag des Kolloquiums:	13. Juni 2007
Dekan:	Prof. Dr. Uli Müller
Vorsitzender:	Prof. Dr. Jörn Walter
Berichterstatter:	Prof. Dr. Volkhard Helms
	Prof. Dr. Hans-Peter Lenhof
Beisitzender:	Dr. Michael Hutter
Protokollantin:	M. Sc. Susanne Eyrisch

Eidesstattliche Erklärung

Hiermit versichere ich an Eides Statt, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichent.
Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, Juni 2007

Sam Ansari

Die vorliegende Arbeit entstand in der Zeit von Februar 2003 bis Dezember 2006 am Zentrum für Bioinformatik an der Universität des Saarlandes unter Leitung von Herrn Prof. Dr. Volkhard Helms.

Zusammenfassung:

Jüngste Forschungsergebnisse der letzten Jahre unterstreichen, dass Protein-Protein Interaktionen eine wichtige Rolle für die biologischen Abläufe im intra- und extrazellulären Raum spielen. Aufgrund der teilweise widersprüchlichen und unvollständigen Erkenntnisse über Protein-Protein Interaktionen, befasst sich die Arbeit mit diesem aktuellen Thema.

In der Arbeit wird zunächst eine große Zahl von temporär wechselwirkenden Proteinkomplexen mit bekannter Struktur gesammelt und analysiert. Dabei werden unter anderem Aspekte wie Vorkommen und Paarungspräferenzen der Aminosäuren und Sekundärstrukturelemente ermittelt, für die in früheren Untersuchungen bereits einige charakteristische Merkmale gefunden wurden. Die Ergebnisse dieses Kapitels stehen im Einklang mit früheren Studien und zeigen, dass temporär wechselwirkende Proteinkomplexe zwecks Reversibilität der Interaktion einen höheren Anteil an hydrophilen Aminosäuren besitzen und eine hohe geometrisch komplementäre Schnittstellenregion aufweisen.

Diese charakteristischen Merkmale werden in einem weiteren Ansatz auf ihre Vorhersagekraft untersucht. Dazu wird die Effizienz von Protein-Protein Dockingprogrammen unter Berücksichtigung dieser Merkmale ausgewertet.

Eine bekannte Schwäche von Protein-Protein Dockingprogrammen besteht darin, dass die nativen Protein-Protein Komplexe allein über die geometrische Komplementarität an der Schnittstellenregion ermittelt werden. Frühere Studien haben gezeigt, dass diese Vereinfachung in vielen Fällen zu einer großen Zahl von falsch-positiven Ergebnissen führt. Mit der Kenntnis über Paarungspräferenzen von Aminosäuren und Sekundärstrukturelementen werden die Ausgaben des Dockingprogramms nun neu analysiert. Tatsächlich zeigen die Ergebnisse leichte Verbesserungen. Eine detaillierte Analyse der Resultate zeigt allerdings, dass die verwendeten Merkmale zu keiner klaren Erkennung von falsch-positiven Dockingergebnissen führen. Um spezifischere Merkmale zur Erkennung von temporär wechselwirkenden Proteinkomplexen ausfindig zu machen, werden anschließend temporär und permanent wechselwirkende Proteinkomplexe miteinander verglichen. Damit werden die Unterschiede dieser beiden Proteinkomplextypen verdeutlicht und eine spezifische Merkmalsanalyse erleichtert. Auf

Basis eines neuen Datensatzes ergeben sich für Paarungspräferenzen der Aminosäuren und Sekundärstrukturelemente keine klaren Unterscheidungen. Diese Beobachtung lässt sich zum Teil auf den geringen Datensatz zurückführen, was daraufhin zu einer ausgedehnten Datensuche und Datenbank-Konstruktion führt. Eine umfangreiche Literaturrecherche ergab 268 temporär und 266 permanent wechselwirkende Proteinkomplexe. Die für diesen Zweck entwickelte MySQL Datenbank erlaubt eine schnelle und spezifische Auswahl von Proteinkomplexen sowie die Berechnung von unterschiedlichen Potentialen, die z.B. in Dockingprogrammen eingesetzt werden können. Weiterhin lassen sich viele zusätzliche Merkmale der Proteinkomplexe zusammenstellen und ausgeben. Mit dieser großen und schnell zugänglichen Datenmenge wird das Problem der klaren Unterscheidung von temporär und permanent wechselwirkenden Proteinkomplexen wieder aufgegriffen. Der Einsatz eines automatisierten Mustererkennungs-Programms und die Analyse von 10.038 Merkmalen oder 347 Merkmalsgruppen anhand von 534 Proteinkomplexen ergab schließlich eine hohe Genauigkeit der Unterscheidung von temporär und permanent wechselwirkenden Proteinkomplexen. Diese Genauigkeit wird durch die gewichtete Kombination von lediglich vier minderdimensionale Merkmalsgruppen erreicht. Mit dieser hohen Wiedererkennung von temporär wechselwirkenden Proteinkomplexen sollte die Effizienz von Protein-Protein Dockingprogrammen, die bei der Evaluation der Dockinganordnungen allein auf geometrische Schnittstellenkomplementarität beruhen, deutlich erhöht werden. Weiterhin erlaubt die hohe Wiedererkennung der beiden Proteinkomplextypen, neue Daten aus der RCSB PDB automatisch zu klassifizieren, und in der Datenbank abzulegen. Eine größere Datenmenge wird die statistische Aussagekraft der Analysen deutlich erhöhen und eine feinere Aufteilung der Komplextypen erlauben.

Kurzbeschreibung:

Protein-Protein Interaktionen haben in den letzten Jahren sowohl im Bereich der Pharmazie, Medizin, Biologie, als auch im Bereich der Bioinformatik großes Interesse erlangt. In dieser Arbeit werden statistische Daten zu transienten Protein-Protein Interaktionen gesammelt und ausgewertet. Charakteristische Mermale werden in einem weiteren Ansatz auf ihre Vorhersagekraft untersucht. Dazu werden die Ergebnisse aus einem Docking-Programm nach diesen Merkmalen bewertet um natürliche Komplexe von solchen, die lediglich eine hohe geometrische Komplementarität aufweisen, zu unterscheiden. Die Ergebnisse zeigen Verbesserungen, aber dennoch Schwächen in der Vorhersagekraft auf. Um noch spezifischere Merkmale ausfindig zu machen, werden transiente und permanente Komplexe gegeneinander verglichen. Der eingeschränkte Datensatz führt schließlich zu einer ausgedehnten Datensuche und Datenbank-Konstruktion. Diese wird schlussendlich für eine sehr detaillierte Merkmalsanalyse verwendet, die ein automatisiertes Mustererkennungs-Programm verwendet. Mit Hilfe dieses Programmes können sogar Kombinationen von Merkmalen auf ihre Spezifität untersucht werden, die schliesslich zu einer hohen Genauigkeit der Unterscheidung von transienten und permanenten Protein-Protein Interaktionen führt. Eine Kombination von vier Merkmalsgruppen ist dabei ausreichend. Damit können nun Docking-Programme verbessert werden, die zum Zwecke der Rechenzeitreduktion die Auswertung der Komplex-Anordnungen nur auf geometrische Komplementarität beziehen.

Abstract:

In the past years protein-protein interactions have gained a lot of interest in the fields of pharmacy, medicine, biology, and bioinformatics. In this work, statistical information on transient protein-protein interactions are collected and analyzed. Characteristic properties are then evaluated and their predictability estimated. Therefore, the results from a common docking approach are re-evaluated with the collected information to discriminate the native structure from those that simply have a high geometric complementarity at the interface region. The results show that although there is a noticeable improvement of the predictability after applying statistical information, the overall accuracy is still low. To find other more specific properties, transient and permanent complexes were compared to each other. The lack of data leads to an extensive search for more suitable structural data and the development of an extensive database. This database was ultimately used to retrieve a large number of protein properties that were automatically analyzed for their separation precision. A high accuracy was obtained in separating transient and permanent interactions based on the combination of only four properties. Combining this information with common docking approaches based on geometrical complementarity may lead to satisfying sensitivities.

Contents

Organization of the Thesis

Chapter 1 gives a short introduction to protein-protein interactions. Starting with the principles of dimerization and oligomerization and their consequences in biological cells will lead us to the current view of protein-protein interactions. Furthermore, this chapter introduces a number of methods and concepts that were employed and presumed in this thesis.

Chapter 2 presents a statistical approach based on non-redundant transient protein-protein interfaces of known structure. By combining many known aspects of protein-protein interactions, deeper insight in their nature was obtained.

Chapter 3 is based on chapter 2 and presents an application of its observations. Using a rigid-body docking approach, the sensitivity of the docking is intensely tested after applying residue and structure based potentials.

Chapter 4 focuses on the specificity of interface properties for transient protein-protein interfaces. Comparing a number of interface properties from transient protein-protein complexes to permanent protein-protein complexes will reveal the specificity of the given interface properties. This was observed by assessing the separation quality of the tested interface properties into transient and permanent protein-protein interactions.

Chapter 5 introduces a new database that was developed to store a large number of properties collected from transient and permanent protein-protein complexes. By storing these data in a database their accessibility is enhanced and allows more detailed and faster statistical analyses.

Chapter 6 is based on the database presented in chapter 5 and employs a machine learning approach to find protein properties within a large dataset that lead to a clear separation of transient/non-obligate and permanent/obligate interactions. In this chapter combinations of properties are analyzed as well.

Chapter 7 gives an overall outlook for this thesis.

1 Introduction

1.1. Protein-Protein Interactions

Almost the full essential structure and function of biological cells may be referred to proteins. These large and complex molecules demonstrate a great flexibility that allows them to perform a large number of activities essential to life. No other type of biological macromolecule could carry all of the functions that proteins have collected over billions of years of evolution. The characteristic structures of proteins allow particular chemical groups to be placed in specific locations on the three-dimensional structure. This precision allows proteins to act as catalysts (enzymes) for a variety of chemical reactions. Precise placement of chemical groups also allows proteins to play important structural, transport, and regulatory functions in organisms.

1.1.1. The Interaction of Proteins

In biological systems proteins rarely act in isolation but bind other biomolecules to initiate cellular processes. These binding partners are often other proteins, as well as copies of the same protein that form dimers or higher-order oligomers, and may occur in relative isolation and within protein interaction networks and cascades [1][2]. Dimerization or oligomerization may provide several different structural and functional advantages to proteins such as improved stability, control over the accessibility and specificity of active sites, as well as increased complexity. Figure 1 shows an overview of the functional consequences of dimerization and oligomerization.

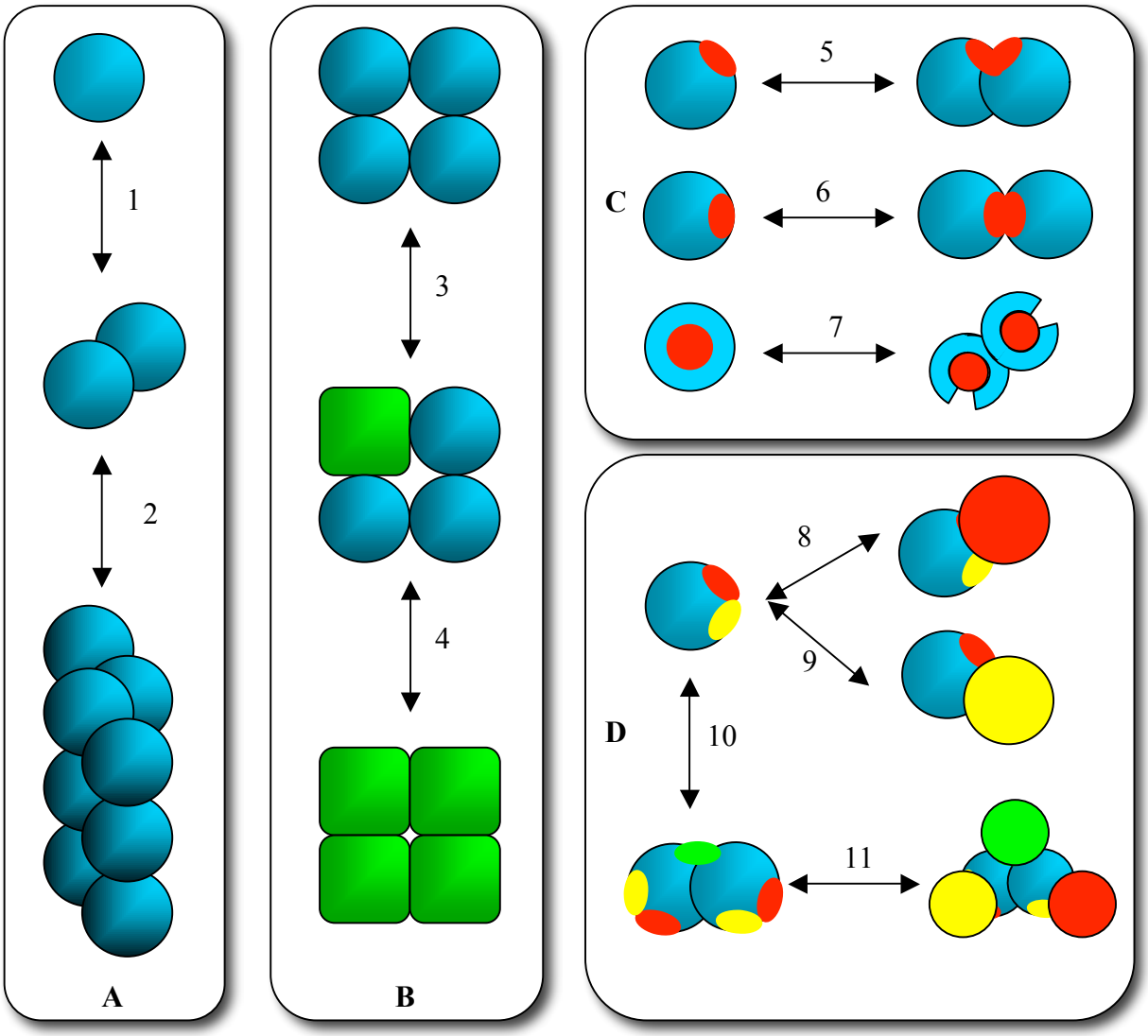


Figure 1: *Functional consequences of dimerization and oligomerization. (**A**) Concentration, stability and assembly. (**B**) Cooperation and allostery. (**C**) Modification of the active site. (**D**) Dimerization yields increased diversity in the formation of regulatory complexes.*

Figure 1A shows the dimerization as a natural process under conditions when the protein concentration is higher than the dissociation constant for the dimerization {1}. The consequence is a higher population of the dimeric form and a lower surface area when compared to the monomers' surfaces. The assembly of large structures from previously formed subunits is a way to form large stable and dynamic structures without increasing the size of the genome or running into problems associated with the folding of large proteins {2}. As shown in figure 1B intramolecular surfaces between monomers and oligomers can generate sites for regulation by allostery. The binding of a cofactor or a

substrate to a single subunit in an oligomer can change the conformation of that subunit {3} and cooperatively induce structural changes in the remaining subunits {4}. Dimerization of a monomer in figure 1C can generate new binding sites at the dimer interface or extend existing binding sites to increase specificity {5}, hide and block {6} or reveal active sites {7}. On the other hand, dimerization can lead to diversity in the formation of regulatory complexes shown in figure 1D. A protein might contain overlapping binding sites for different proteins. In this case the monomer can only bind a single competing protein at the same time {8}{9}. However, dimerization may also enable the simultaneous binding of those proteins on different subunits and create new binding sites for additional proteins {10}{11}.

1.1.2. The Role of Protein Dimerization and Oligomerization in Biological Cells

One of the major problems in understanding the role of protein dimerization in biological cells lies in the rather small amount of available biophysical data when compared to the numbers of known proteins. The best-characterized protein class is certainly the class of enzymes. Several different factors were proposed to explain the large frequency of occurrence of dimerizing and mainly oligomerizing enzymes. Multimeric enzymes mostly form their active sites at the subunit interface, which leads to a high local concentration of active sites. The consequence is an enhanced regulation with loss in enzyme activity. In detail, the generation of new intermolecular interfaces can produce sites for allosteric regulation, enabling cofactors to bind to nonsubstrate sites, or facilitating substrate-induced cooperation. Hemoglobin is a classic example of a protein complex undergoing structural changes upon ligand binding together with the corresponding generation of a conformation with a very high ligand-binding affinity (figure 2). In lower vertebrates such as snakes, oxygenation causes dissociation of the hemoglobin tetramer to produce a dimer that acts as an oxygen store because of the higher affinity of the dimer for oxygen [3]. Under certain conditions such as stress or high activity, an associated decrease in pH promotes ATP-induced tetramerization and allosterism, which then results in the release of oxygen. This transition from dimer to tetramer conformation might represent an intermediate point

during the evolution of the more stable hemoglobin tetramer that is found in higher vertebrates, where cooperative ligand binding is based on switching between quaternary states with different affinities for oxygen.

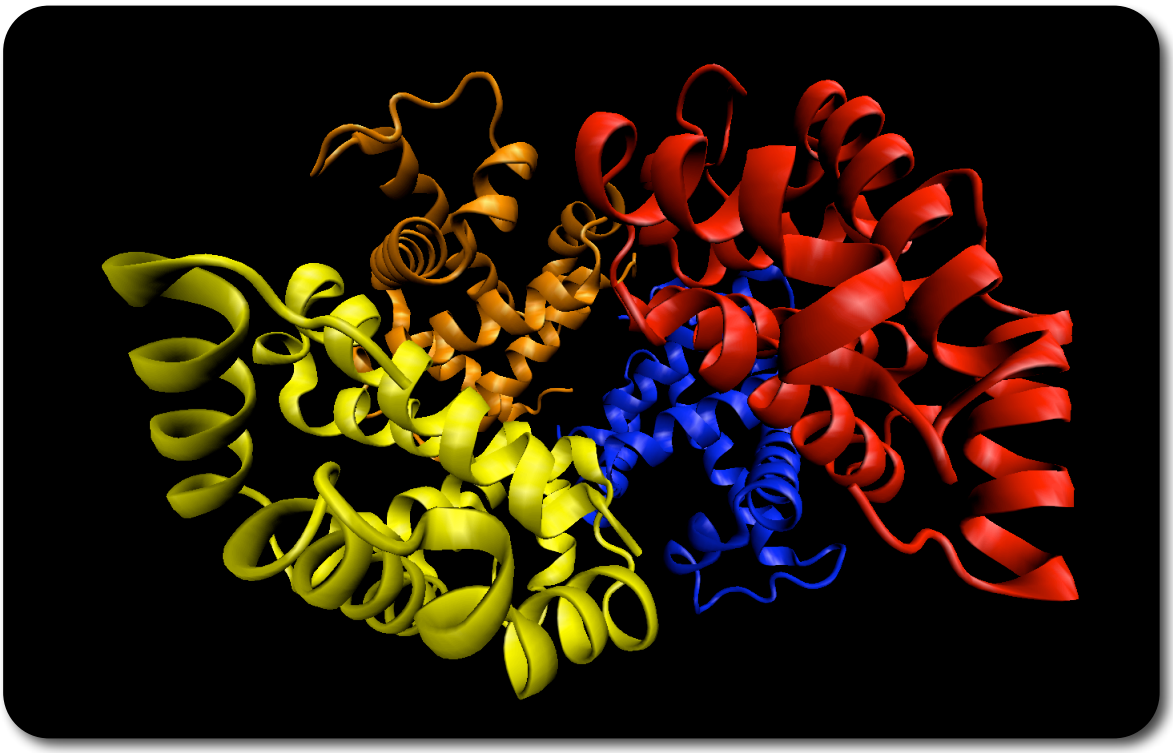


Figure 2: *Human Hemoglobin A tetramer. It is composed of four protein chains, two α-chains (red and blue) and two β-chains (brown and yellow), each with a ring-like heme group containing an iron atom (not shown). Picture rendered in VMD using the PDB 1buw [4].*

Another motive for the dimerization is the mechanism for enzyme activation. As an example for this type, the family of cysteine and aspartic acid-proteases (caspases) will be discussed. Caspases are single-chain enzymes controlling the process that leads to cell death during apoptosis, which is one of the main types of programmed cell death. Failure of apoptosis mostly contributes to tumor development and autoimmune diseases. In general, limited proteolysis of the caspase generates two active catalytic domains, which is quite common for many other protease activations. However, this mechanism alone cannot lead to the activation of the initial protease (caspase-9) in the caspase pathway (figure 3). This is because there is no activating protease upstream of this enzyme. Structural and experimental studies have shown that under physiological conditions caspase-9 exists as an inactive monomer. During apoptosis, the cofactor Apaf-1 and

caspase-9 form a 1:1 complex in the presence of cytochrome *c* and ATP and generate an apoptosome [5]. This oligomeric complex colocalizes with multiple caspase-9 molecules causing an increase of the local concentration of caspase and inducing dimer formation and activation of the enzyme by exceeding the dissociation constant K_d for homodimerization. The interface of the dimer is formed by the interaction of an exposed activation loop in one monomer unit with a hydrophobic pocket in the other monomer. This interaction stabilizes the priming bulge of the activation loop and enables the active site that forms the substrate-binding conformation. Caspase-9 can then provide the proteolytic activity generating the active forms of caspase-3 and caspase-7.

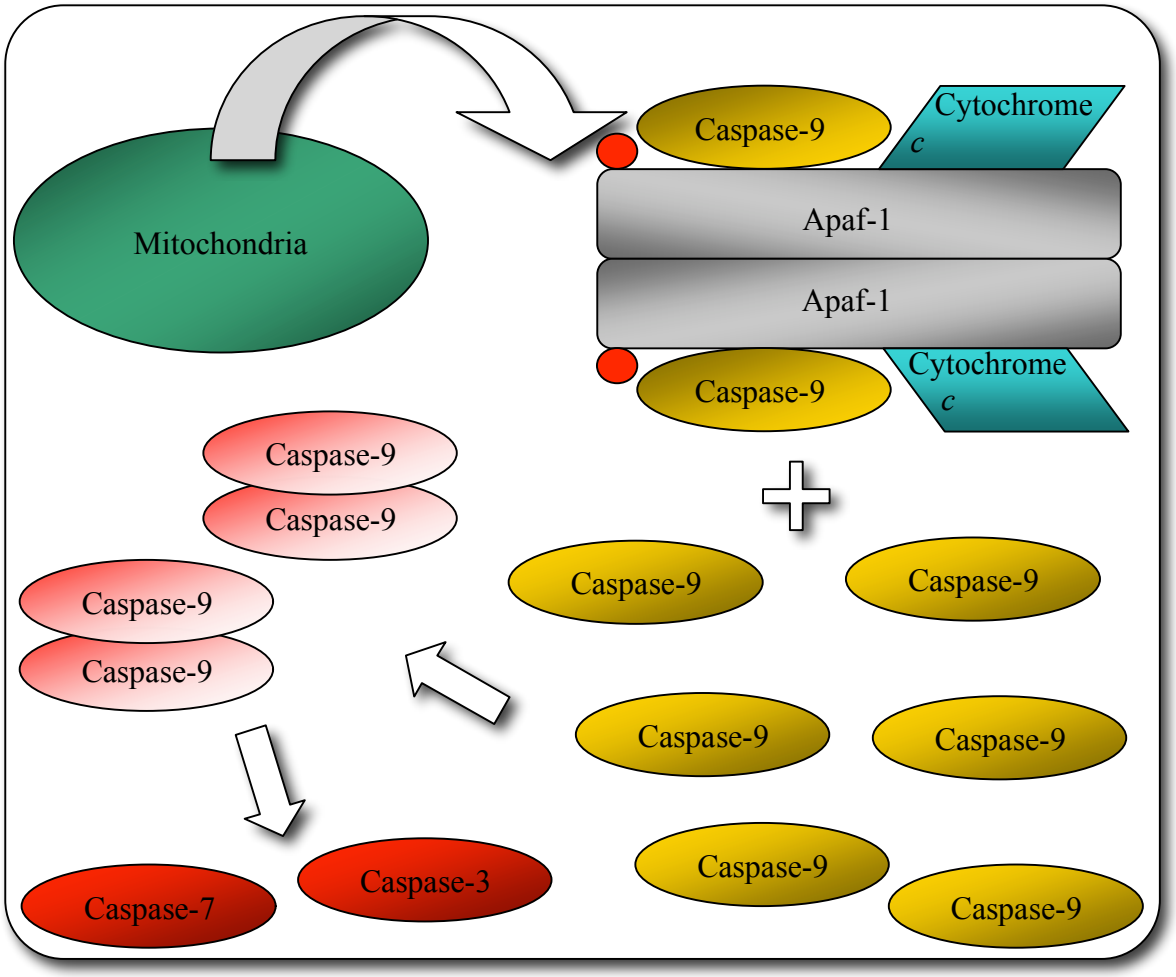


Figure 3: *Caspase pathway. The mitochondrial stress causes a release of cytochrome c (blue) from mitochondria (green), which then interacts with Apaf-1 (grey), ATP (red ball), and the inactive form of caspase-9 (dark yellow) forming a dimer. In the presence of inactive caspase-9 monomers (yellow), this complex induces caspase-9 dimerization and activation (light red). These activated caspases-9 further proceed to activate the effector caspases (red) that initiate the cleavage of various cellular targets.*

On the other hand, dimerization can also inhibit an active monomeric enzyme as the receptor-like protein tyrosine phosphatase-α (figure 4), which mostly exists on the cell surface as a weak homodimer. The activity of this homodimer is down-regulated because of a part of one monomer that is stuck into the active site of the other monomer [6]. Such dimers could be activated with the binding of ligands that favor the dissociation to active monomers, or by signaling events that induce an open conformation in the dimer (figure 5).

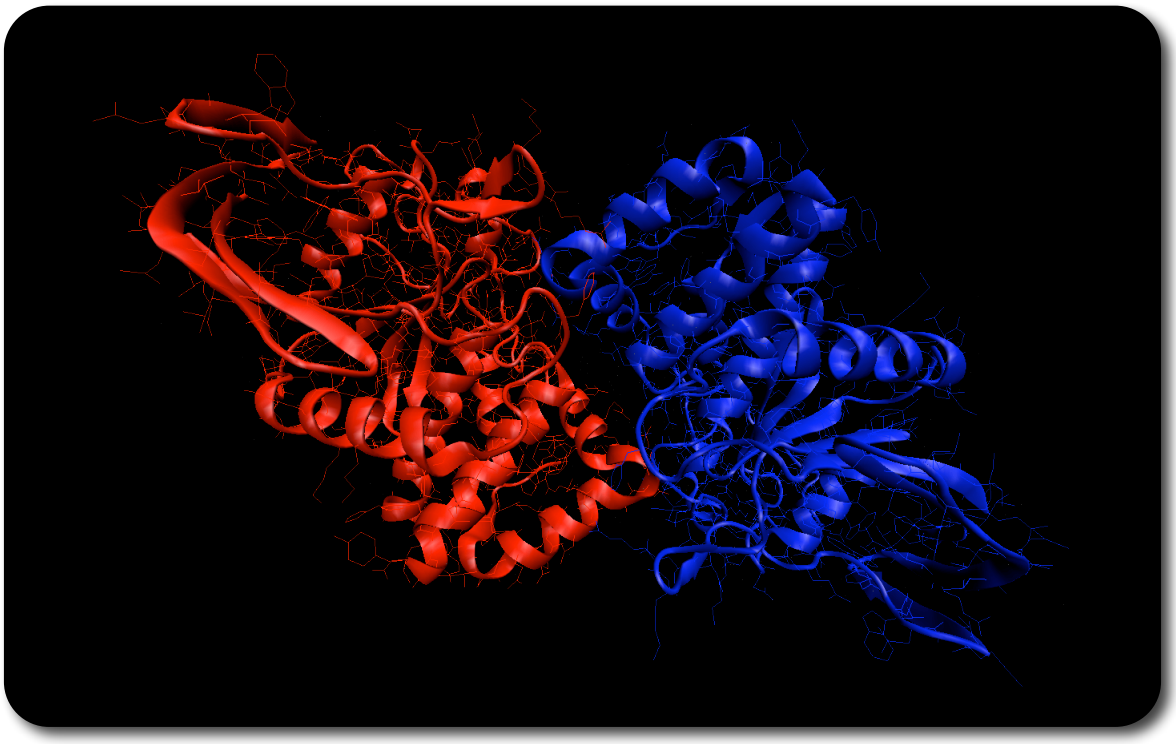


Figure 4: *Domain 1 of the receptor-like protein tyrosine phosphatase-α of mouse. Chain A (red) and chain B (blue) dimer in the inactive state. Picture rendered in VMD using the PDB 1yfo [7].*

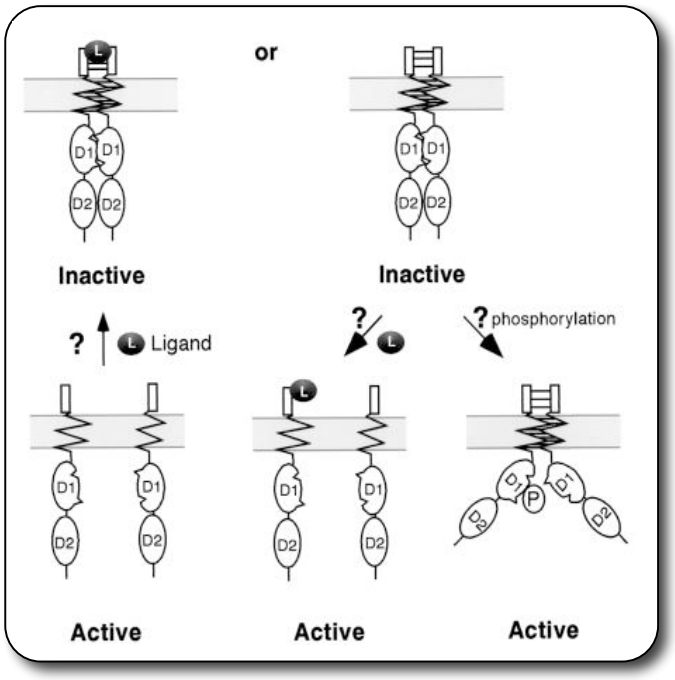


Figure 5: *Model for the regulation of receptor-like protein tyrosine phosphatase-α by dimerization. In the inactive state, the receptor-like protein tyrosine phosphatases-α are dimerized via domain D1 (also see figure 4), the transmembrane domain, and the extracellular domain. In the active state, the receptors are either monomers or dimers that no longer dimerize via D1 due to phosphorylation. Ligand binding can either stabilize or destabilize dimers [6].*

Cell-surface receptor oligomerization and activation in response to the binding of an agonist is a common theme in the pathways transfering a signal across the cell membrane. Examples are the receptor families of growth hormones, interferons, cytokines and tyrosine kinases [8]. G-protein-coupled receptors (GPCRs), which are the most common cell-surface receptors, mostly function as dimers [9]. For some receptors agonist binding is required for initiation of the dimerization, while others require the homodimerization before the agonist can bind. Additionally, it was shown that rhodopsin must be arranged in dimeric arrays to absorb single photons [10]. Although dimerization can involve covalent interactions [11], most GPCRs dimerize via non-covalent interactions between extracellular domains, transmembrane regions and C-terminal tails of the proteins [12].

As previously mentioned, the oligomerization of multiple and identical subunits provides a simple way to form large structures. Structures such as the long fibrous extracellular matrix proteins myosin and collagen can be very stable and can last a lifetime (figure 6). On the other hand, some are rather dynamic, e.g. tubulin heterodimers that are composed of α and β subunits. These subunits can be added or removed from the end of microtubuli to form the structural and transport system of the cytoskeleton [13] (figure 7).

Figure 6: *Strong complex between the I domain of integrin α2β1 (blue) and a triple helical collagen peptide (red, grey, orange). Picture rendered in VMD using the PDB 1dzi [14].*

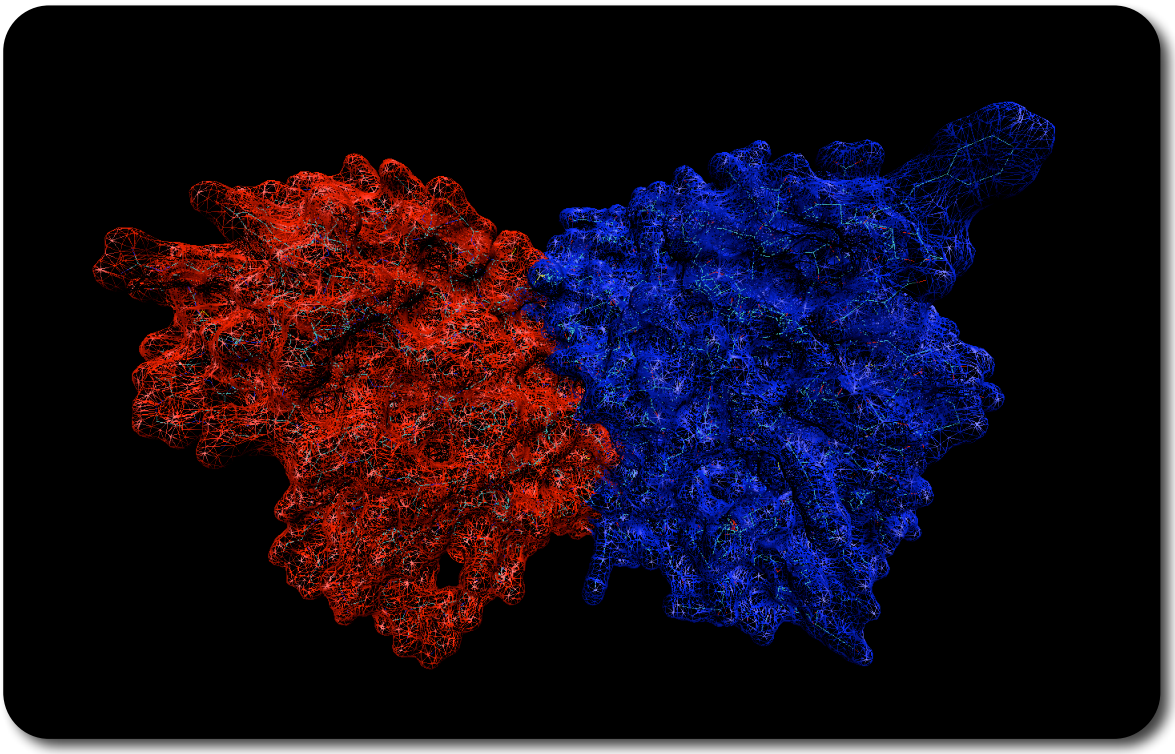


Figure 7: *Dynamic complex of an α-β tubulin dimer (red, blue). Picture rendered in VMD using the PDB 1tub [15].*

1.1.3. The Diversity of Protein-Protein Interactions

In their review on the diversity of protein-protein interactions in 2003, Nooren and Thornton discussed the structural and functional diversity of protein-protein interactions primarily based on protein families with available structural data [16]. As mentioned in the previous section, the protomers' localization, concentration and local environment affect the interactions with the same or other protomers. In their review, the authors laid the basis for most subsequent studies on protein-protein interactions that considered the differentiation of protein-protein interaction types. Nooren and Thornton specified three classes of complexes: homo/hetero-oligomeric complexes, non-obligate/obligate complexes, and transient/permanent complexes. Homo/hetero-oligomeric complexes are interactions between identical or non-identical chains. In general, homo-oligomers can have either an isologous or heterologous organization. Isologously organized associations lead to the same surface of the two monomers (figure 8A) while heterologous assemblies use different interfaces and lead to different surfaces (figure 8B). Such heterologous assemblies can form further oligomerizations.

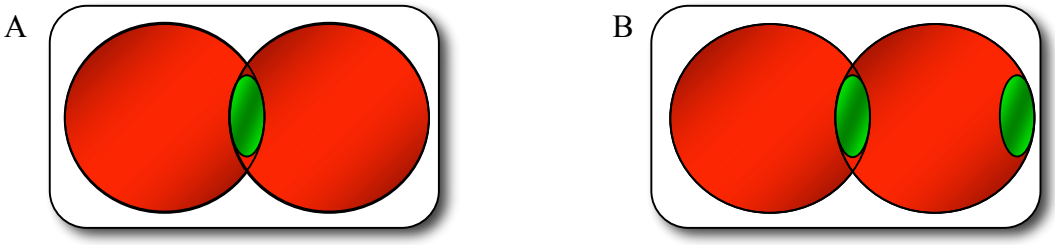


Figure 8*:(A) Isologous and (B) heterologous homo-oligomers.*

The class of non-obligate and obligate complexes is defined on the basis of whether a complex is composed of protomers that are found or not found as stable *in vivo* structures. Complexes that do not have stable unbound protomers are called obligate complexes since their bound form is required and obligated. Non-obligate complexes are formed from stable unbound protomers that form in a dynamic equilibrium between unbound and bound states. This is the case for many intracellular signaling complexes and enzyme inhibitor complexes (e.g. figure 7). The last class mentioned in the review of Nooren and Thornton is the group of transient and permanent complexes. Transient complexes are believed not to form a complexed state for the whole lifetime of the

protomers. Such interactions are not stable enough to simply last in the complexed state. They associate and dissociate *in vivo*. Opposite, permanent interactions are stable and mostly remain in the complexed state for the lifetime of the protomers. Furthermore, the authors distinguish between weak and strong transient interactions. While weak transient interactions exist in a dynamic oligomeric equilibrium in solution, strong transient interactions require a molecular trigger to shift the oligomeric equilibrium. However, many protein-protein interactions cannot be clearly separated into either of the two last classes. The stability of the unbound protomers is rather relative and strongly depends on the physiological conditions of the environment. Additionally, it should be noticed that the two classes of non-obligate/obligate and transient/permanent complexes are very closely related. Obligate complexes that do not have a stable dissociated form mostly remain in the complexed state for their entire lifetime – also covered by the specification of permanent complexes. Transient interactions that dissociate and associate require stable unbound protomers like non-obligate complexes. In the literature these two classes are mostly combined. Although they focus on different aspects of complexes, a separation of protein complexes into these two classes will most likely lead to the same distribution. Remarkably, this is not the case for antigen-antibody complexes. Although antigens and antibodies do occur in a stable structure in solution such as non-obligate complexes, their complexed state has a strong binding, as it is the case for permanent complexes.

At the same time, Ofran and Rost published a study where they classified protein-protein interactions into six different types of interfaces [17]. By introducing a new data-mining method the authors differentiated protein interfaces into:

1. intra-domain interfaces that are within one structural domain;
2. domain-domain interfaces that occur between different domains within one chain;
3. homo-obligomer interfaces that form between permanently interacting identical chains;
4. homo-complex interfaces formed between transiently interacting identical protein chains;
5. hetero-obligomer interfaces formed between permanently interacting different protein chains;
6. hetero-complex interfaces which form associations between different transiently interacting protein chains.

The authors introduced the term "obligomer" that stands for obligate oligomers, where "complex" stands for non-obligate oligomers. Their definition for obligate and non-obligate complexes is similar to Nooren and Thorntons' as obligate/non-obligate are also called two-state/three-state complexes. Folding and binding of the interacting proteins are inseparable for two-state complexes. Thus, such interactions form a permanent complexed state. On the other hand, protomers of non-obligate complexes fold independently and then bind. Such complexes are also called three-state complexes [18]. Since a change in quaternary state is often coupled with biological function or activity, three-state or transient/non-obligate protein-protein interactions are important biological regulators and are particularly emphasized in this work.

1.1.4. Known Properties of Protein-Protein Interactions

Many previous studies analyzed the properties of protein-protein interactions. Since the number of three-dimensional structural data available in earlier years was rather limited, the initial studies mostly examined general properties of interfaces such as the size of the contact area, the polarity of the interface, protrusion and flatness [19][20][21][22][23]. These and many other studies form the basis of the current understanding of protein-protein interactions: interfaces of obligate complexes that are mostly formed by homodimers are larger and more hydrophobic than non-obligate associations [24][25]. The stable association derives from the co-folded and co-expressed protomers and the large hydrophobic surface patches, which are causing strong and tight interactions. In contrast, non-obligate interactions rather exhibit a more polar interface ensuring the stable unbound state of the monomers. LoConte et al. furthermore noticed conformational changes of the protomers upon complex formation once the interface area is larger than 1000Å^2 [25]. The consequences of such conformational changes may lead to an induced-fit which increases the lifetime of an interaction. However, although some structural differences were found between obligate and non-obligate complexes, the difficulty still remains to efficiently separate their protein-protein interactions. There exists a continuum between non-obligate/obligate or transient/permanent interactions and previously

mentioned structural characterization properties appear inadequate to distinguish between their different affinities or specificities.

One important property of protein-protein interactions is obviously the specificity of interaction. Due to the rather crowded environment *in vivo,* many protomers are not in direct vicinity and need to be highly specific in partner recognition and binding, as it is the case for hormone-receptor and enzyme-inhibitor complexes. Such specific interactions mostly form interfaces with strong geometric and chemical complementarity. However, there are also multispecific interactions where multiple binding partners compete. Such complexes are mostly co-localized and their specificity is rather low.

With an increasing number of available data the functional and structural principles of protein-protein interactions and their great diversity may soon be thoroughly understood.

1.2. Methods in Bioinformatics

In this section, a number of basic methods and techniques that were employed in this work will be briefly introduced. Describing the concept of databases and their typical implementations will eventually lead to the concepts of sequence comparisons. These comparisons are based on the concept of molecular evolution and will be briefly described as well. Furthermore, several structural analysis methods will be mentioned, which will lead to the last topic: data mining.

1.2.1. Databases

A database is a collection of information that is systematically stored in a computer and can be accessed with querying the dataset and consulting it to answer questions. There are two main motivations for storing data on a computer: retrieval and discovery. Retrieval is basically the ability to access stored data. The growing number of sequence information would be useless in its essence if there were no possibilities to retrieve the data. However, it is even more important to retrieve additional knowledge from the system than what was stored. Such additional information can be obtained with detection of connections between two pieces of information that were not known to be related at

the time they were separately stored in the database. Another way is to perform computational approaches on the data, which may yield new insight into the records.
In this section two major groups of databases are separately described: sequence and structural databases. While sequence data contains just sequences of the proteins or nucleotides, their rich annotations and large number in the databases makes them essential for further analyses. Structural data is more preferential for most analyses but due to the difficulties in generating such data, there are a number of computational efforts to overcome this restriction as described in section 1.2.1.2.

1.2.1.1. Sequence Databases

Among all available databases the largest are without doubt sequence databases. A sequence database is mostly a collection of nucleotides or amino acids containing data from specific organisms or all. Currently, nucleotide sequence databases with up to 80 million entries mark the largest amount of data in a database [26]. This is due to great success in recent international genome projects. However, maintaining databases is a great challenge. The major problem arises when joining records from a wide range of sources and individual researchers. The sequences and especially the biological annotations attached may qualitatively vary. There is also much redundancy, as multiple labs often submit numerous sequences that are nearly identical to other available entries. Another issue is based on the way sequences are retrieved. Protein sequence databases are mostly based on automated translations of mRNA nucleotide sequences, where all six open reading frames (ORFs) are considered and the meaningful ORF is translated and stored. This method is very appropriate when compared to other costly and time-consuming methods such as mass spectrometry and the Edman degradation reaction. However, this automated approach barely leads to qualitatively competitive annotations and requires semi manual modifications, which lead to the large number of available and different protein sequence databases.
Based on the primary sequence information a large number of secondary databases arose by time. Secondary databases collect data from primary databases that store annotations and sequences and use certain classification rules to group these sequences. In most cases, functionally or evolutionally related proteins are grouped into one class and their

sequence pattern is then retrieved and used for identifying other yet unknown sequences (also see section 1.2.4). Following, one popular representative for nucleotide and protein sequence databases, and one example for a secondary database are described.

1.2.1.1.1. EMBL (Release) [27]

The EMBL Nucleotide Sequence Database at the EMBL European Bioinformatics Institute (EBI) offers a large set of publicly available nucleotide sequences and annotations. Collaborations with DDBJ [28] and GenBank [29] led to coverage of the whole genome sequencing project data. The most common technique is expressed sequence tag (EST). EST is a short sub-sequence of a transcribed protein coding or non-coding nucleotide sequence. It was originally used to identify gene transcripts, but has become a common method in gene discovery and sequence determination. The whole genome shotgun (WGS) sequencing is a faster and more complex sequencing process when compared to the common chain termination method of DNA sequencing after Sanger [30] that can only be used for short strands and makes it necessary to divide longer sequences up and then assemble the results to retrieve the overall sequence. In WGS sequencing, DNA is sliced randomly into small segments, which are then sequenced using the common chain termination method. Multiple overlapping segments for the target DNA are obtained with performing several fragmentation and sequencing rounds. These overlapping segments are then computationally assembled into a contiguous sequence. Although WGS sequencing is available for many years now, it became preferential when Celera Genomics announced using this method to produce a draft human genome sequence faster than the publicly funded Human Genome Project. Table 1 shows the distribution of the number of submitted sequences and their retrieval technique.

Currently, EMBL Release consists of more than 80 million entries. Table 2 shows the number of entries for some organisms where plants, humans and other mammals stand in the major focus of these genome projects.

Class	Number of Entries
Constructed	841,474
Expressed Sequence Tag	38,355,718
Genome Sequence Scan	15,345,539
High Throughput cDNA sequencing	440,827
High Throughput Genome sequencing	94,210
Patents	3,404,841
Standard	3,186,797
Sequence Tagged Site	883,330
Third Party Annotation	5,119
Whole Genome Shotgun	18,034,036

Table 1: *Distribution of the submitted sequences and their retrieval technique for nucleotide sequences in EMBL [31].*

Division	Number of entries
Environmental Samples	1,732,858
Fungi	1,392,128
Human	11,448,482
Invertebrates	9,640,399
Other Mammals	16,973,288
Mus musculus	8,160,933
Bacteriophage	3,936
Plants	17,893,858
Prokaryotes	493,678
Rodents	3,607,057
Synthetic	678,974
Unclassified	1,203,436
Viruses	403,338
Other Vertebrates	6,959,526

Table 2: *Distribution of the submitted nucleotide sequences and their organisms in EMBL [31].*

1.2.1.1.2. UniProtKB/SWISSProt [32]

The SWISSProt database is a popular primary sequence database containing protein sequences. Just recently it was renamed to UniProtKB (Universal Protein resource KnowledgeBase). Although the SWISSProt database with only quarter million entries is

one of the smallest protein sequence databases, it is yet the most popular protein database. This is due to its rich and partial manual annotations. Each entry contains the core data – sequence data, bibliographical references and taxonomic data – and annotations describing the function of the protein, post-translational modifications, domains and sites, secondary structure and quarternary structure, similarities to other proteins, diseases associated with deficiencies in the protein, sequence conflicts, and more. However, the developers also focus on crosslinking the entries with those of other databases with different contents, e.g. the nucleotide sequence database EMBL (Release), the protein structure database PDB, and various protein domain and family characterization databases (PRINTS, Pfam, INTERPRO, and more). At the moment, there are up to 60 references to other databases. SWISSProt is based on the data collected in the TREMBL database which stands for translated EMBL. Automatic translations and simple annotations from the nucleotide sequences in EMBL (Release) are first stored in TREMBL. Manually revising the entries of TREMBL leads to the SWISSProt database. TREMBL is currently more than 10 times bigger than SWISSProt although the database is growing faster (figure 9). Figure 10 shows the distribution of the data by organisms. Interestingly, the distribution is not clearly correlated to that of the nucleotide sequence database EMBL. The focus of the semi-manual annotation lies much stronger on prokaryotic data when compared to the focus of nucleotide data (table 2).

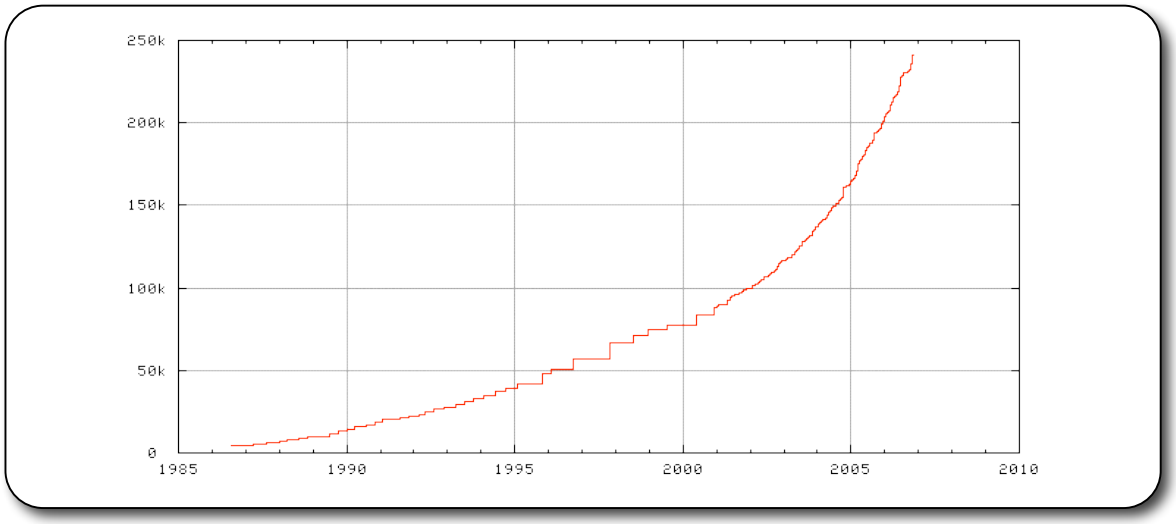


Figure 9: *Number of entries in UniProtKB/SWISSProt in 1000 (k) at various times [33].*

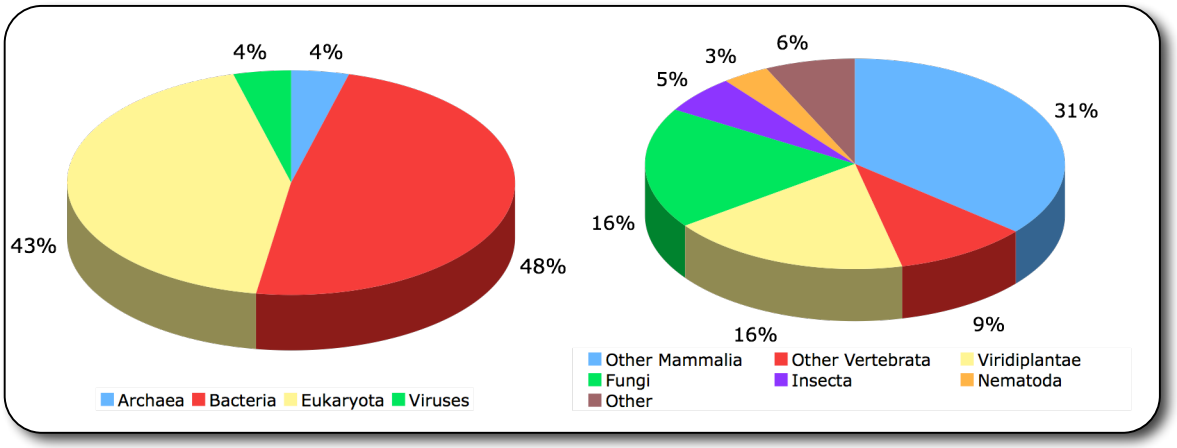


Figure 10: *Taxonomic distribution of the sequences in UniProtKB/SWISSProt. The left graphic shows all organisms and the right graphic only the distribution within eukaryotic organisms [33].*

1.2.1.1.3. COG [34]

Secondary databases were previously presented containing classifications of sequences after a given feature. The COG (Cluster of Orthologous Groups) clusters proteins that are assumed to have evolved from an ancestral protein. Such proteins can be either orthologs or paralogs (also see section 1.2.2). Orthologous proteins stem from different species that diverged from a common ancestor and typically kept the same function. A COG is derived from comparing each protein sequence against all other sequences encoded in completely sequenced genomes. Considering a protein from a given genome, this comparison would reveal those proteins from each of the other genomes to which it is most similar. The relation is tested inversely. If a reciprocal best-hit relationship between these proteins can be found then those that are reciprocal best hits will form a COG. 66 genomes are currently included in the database covering more than 70% of all protein sequences in these genomes. Due to the expected functional relation of COG members to each other, the COG is a well-known sequence analysis database for finding functionally relevant regions.

1.2.1.2. Structural Databases

Understanding of functional and structural principles of protein folding and protein binding is crucially based on the knowledge on three-dimensional protein structures. This is not only due to higher conservation of the protein structure when compared to the proteins' sequence. Several large sequence databases significantly contributing to the current understanding of protein function were mentioned before. However, without any available structural data, sequence data probably would not have led to as much knowledge. The combination of the rather small number of available structural data and the availability of a large number of sequence data reduces the problems caused by the lack of essential structural data. The gap between available structural data and sequence data is certainly related to the methodical difficulties to retrieve structural data. At the moment, structures are typically obtained by X-ray crystallography (table 3). The technique of X-ray crystallography records and analyzes data from the diffraction of X-ray photons arising from their interactions with the electrons of the sample. This generally allows determining type and positions of heavy atoms in a crystallographic lattice. The basis and also most challenging part of this assessment lies in generating crystals of the molecules. An alternative structure determination method is NMR spectroscopy. Here, the sample is first prepared then resonances are assigned, restraints are generated and a structure is calculated and validated. This technique is limited to small proteins due to overlapping peaks in larger proteins and faster weakening magnetization, which leaves less time to detect the signal.

These two methods have complementary features. X-ray crystallography represents a robust and fast approach for proteins that form suitable crystals. NMR has advantages for structural studies of small proteins that are partially disordered, exist in multiple stable conformations in solution, or do not crystallize easily. NMR spectroscopy is an incremental method that can rapidly provide useful information concerning overall protein folding, local dynamics, existence of multiply-folded conformations, or protein-ligand or protein-protein interactions.

The creation of images of molecular structures is one of the most simple and broad applications. Other opportunities arising from structural information are classifications where similar structures are clustered together in order to form families of proteins (secondary databases). As previously mentioned, lot of focus was put in filling the gap

between available structural and sequence data as previously mentioned. Applications such as homology modeling construct a model of a proteins' tertiary structure based on its amino acid-sequence [35][36][37][38][39][40][41]. This technique relies on a sequence alignment between the sequence of unknown structure and at least one related sequence of which the structure could be determined experimentally. Since protein structures are stronger conserved than protein sequences, sequence similarity usually implies significant structural similarity.

	Proteins	NA	Protein/NA complexes	Other	Total
X-ray diffraction	30,746	931	1,421	28	33,126
NMR	4,853	726	122	6	5,707
Electron microscopy	91	10	33	0	134
Other	77	4	3	0	84
Total	35,767	1,671	1,579	34	39,051

Table 3: Structures contained in *PDB on 09/26/2006. NA stands for Nucleic Acid [42].*

1.2.1.2.1. RCSB PDB [43]

Currently, the standard depository for information about the three-dimensional structures of large biological molecules is the RCSB PDB. Founded in 1971 by Brookhaven National Laboratory, management of the Protein Data Bank was transferred in 1998 to members of the Research Collaboratory for Structural Bioinformatics (RCSB).

The PDB format consists of a collection of fixed format records that describe the atomic coordinates, chemical and biochemical features, experimental details of the structure determination, and some structural features such as secondary structure assignments, hydrogen bonding, and biological assemblies and active sites. A large number of databases and projects were developed to integrate and classify the PDB in terms of protein structure, protein function and protein evolution.

At the moment, there are nearly 40,000 structures stored. Compared to the number of known protein sequences of nearly 7,900,000 [44], 40,000 structures seem quite few. This difference mainly arises from the techniques of generating protein structures as discussed in the previous paragraph. However, as figure 11 shows, the number of determined structures in a year increases.

The web presence of the RCSB PDB contains a powerful database interface. The full content of the PDB files can be queried by many properties and features of the entries such as type of chains, number of chains, chain length, header descriptions, enzyme class numbers, and more. Given the importance of non-redundant data, the RCSB PDB site also optionally performs a redundancy assessment on the search results based on sequence alignments and their level of identity (also see section 1.2.3).

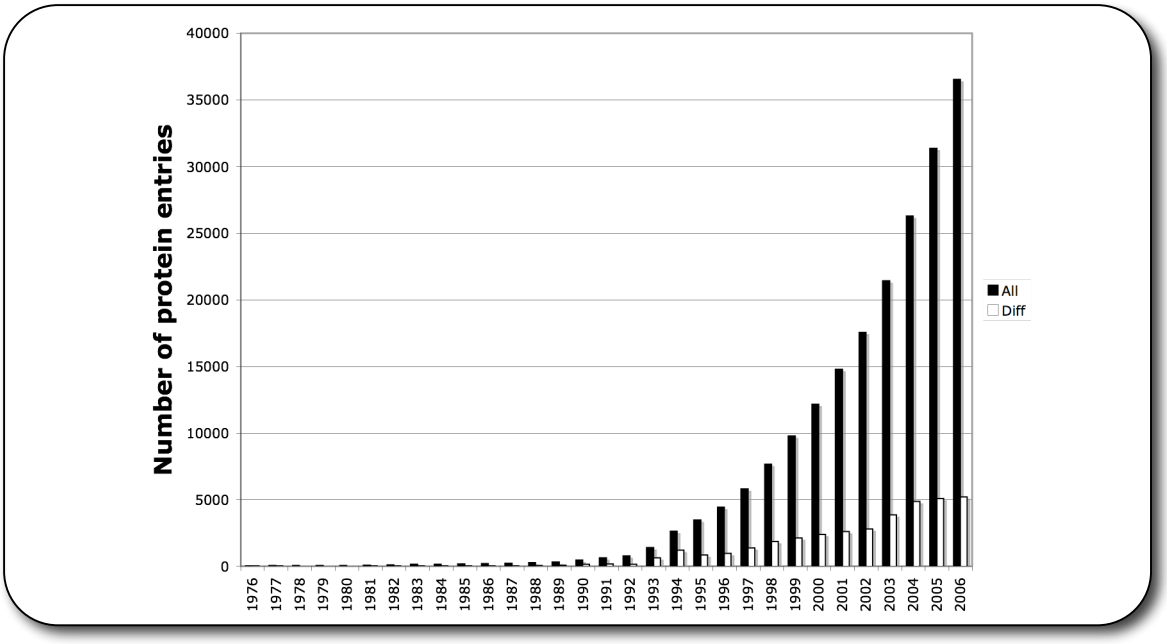


Figure 11*: Yearly growth of protein structures. 'All' are the fully available structures at a given year and 'Diff' shows the increased number of entries compared to the previous year [42].*

1.2.1.2.2. CATH [45] and SCOP [46]

The most common secondary databases of protein structures are CATH (Class, Architecture, Topology, Homologous superfamily) and SCOP (Structural Classifiction Of Proteins). Both databases cover almost the full PDB content. Classifying protein structures invokes separating them into groups in a way that they have similar attributes, such as secondary structure element-composition and other structural attributes. Considering that protein sequences can be grouped into evolutionary families and the fact that protein structures are more strongly conserved than protein sequences, the classification of proteins by structural criteria suggests to be more accurate than based on

sequence similarity or even homology. CATH and SCOP both hierarchically classify the protein structures from the PDB and have similar classes. Although their employment seems essential, one should note that many proteins with the same fold have emerged by divergent evolution from a common ancestor. However, it is also equally possible that they have no common ancestor and adopt the same fold simply because that fold is favorable from a physicochemical point of view.

1.2.2. Molecular Evolution

Understanding the nature of the machinery in organisms is mostly utilized in understanding their development, here evolution. In biology, evolution means the change in heritable attributes of a population over successive generations. All organisms on earth are related to each other through a common ancestor, which makes evolution the source of the vast diversity of organisms on earth. Darwins' theory of evolution divided the procedure of evolution into three major features repeating in an endless cycle:
1. Generation of random variations; 2. Natural selection of the variations; 3. Differential reproductive success. The generation of random variations happens mostly at the level of the genotype, which considers changes within the DNA sequence. Such variations are caused by mutations and insertions/deletions of nucleotides in the sequence. Mutations are mostly errors caused by the DNA replication or DNA repair machinery and provide the genetic variation upon which natural selection can act.
Since most genetic mutations happen at the genotype level they can be neutral in their phenotypic effects or deleterious where they are removed by negative selection. Rarely, mutations may lead to an advantage such as a survival and reproductive advantage, which then may pass on more copies of their genetic material due to their large number of offspring. This is also called positive selection. The accumulation of small changes can result in the evolution of DNA or RNA sequences with new associated phenotypic effects. This process also leads to the evolution of entirely new biological functions.
Homology describes the evolutionary relationship of sequences or structures that diverged from a common ancestor. Since the observation of this relation can only be inferred from sequence or structural similarity, its application is not trivial. Namely, defining threshold

values for similarity is the most challenging part. On the other hand, analogy refers to adopting a shared feature such as protein fold or function by convergent evolution from different ancestors. Given homologous sequences or structures it is possible to separate those that have resulted from gene duplication events within a species genome and perform different but related functions within the same organism (paralogs) from those that perform the same or a highly similar function in different species (orthologs).

1.2.3. Sequence Analyses

7,9 million protein sequences [44] are currently stored in the largest protein sequence database. Given the rich availability of protein sequence data it becomes important to understand how these proteins function. Experimentally characterizing their biochemical properties is impractical. However, since proteins with similar sequences have diverged from a common ancestral gene and possess in most of the cases similar structures and functions [47][48], the development of reliable sequence comparison methods was one of the major foci in bioinformatics.

As mentioned before, protein structure is more conserved than protein sequences. Therefore, structural data is more suitable for analyzing evolutionary relationships. However, due to the lack of structural data as well as rather difficult structure comparison approaches, computational analyses of any newly determined protein sequence typically involve comparing that sequence against libraries of sequences to find related proteins with known functional properties.

The basis of these computational methods requires estimating evolutionary events between two homologous sequences. The common concept is: consider any evolutionary event and weight its probability with scores and penalties. Minimizing these events should most likely lead to the best alignment between two sequences. In most cases, these methods have almost no computational weaknesses. However, the alignment output is strongly dependent on the scores and penalties for evolutionary events. Specifying these scores is mainly based on empirical studies where a number of related sequences are analyzed and evolutionary events statistically evaluated. Based on different datasets there are a number of available scoring matrices for amino acid-exchanges [49][50].

Analyses within protein families have revealed significant changes in the sequences of related proteins as long as they do not affect the folding or stability of the protein [51][52][53]. Other studies have shown that sequences of 100 residues or more, sharing at least 35% identical residues, are most likely homologs [54]. Even at lower levels of sequence identity, where functional annotation is not certain, sequence alignment enables the identification of equivalent regions or residues that may be functionally important. Particularly, multiple sequence alignments that optimize the alignment of several homologs (see 1.2.3.2) can be used to search for patterns of highly conserved residue positions. In this case, pairwise sequence alignment methods are performed to detect close homologs (≥35% identity) and to reveal evolutionary relationships in what Doolittle has defined as a twilight zone of sequence similarity [47] down to as low as 25% identity. Below that, multiple alignment methods must be used to infer homology.

1.2.3.1. Pairwise Protein Alignments

The methods for comparing protein sequences can be divided into fast approximate approaches and those that attempt to accurately determine all possible residue positions. Fast approximate methods are mostly used for scanning a database with a sequence of unknown notation in order to find a homolog of known notation. Any relatives identified in the database can then be realigned using the accurate, but slower, methods.

Pairwise alignments find their origin in 1970. At that time Needleman and Wunsch presented an algorithm for efficient comparison of two protein sequences [55]. By dividing the alignment into sub-alignments the comparison performed reasonably fast. Any possible orientation of the alignment is evaluated. Today the Needleman & Wunsch algorithm is known as a dynamic global alignment, where all possible alignments along the entire sequences are evaluated and accurately determined. However, many proteins are modular and comprise more than one domain. Domain recruitment and domain shuffling are now established as very common evolutionary mechanisms with which organisms expand their functional repertoire. Because of this, proteins that share one or more homologous protein domains may not be homologous over their entire sequence length. Therefore, 11 years after Needleman and Wunsch, Smith and Waterman developed a local implementation of the dynamic programming algorithm which seeks a local region of similarity [56].

Dynamic algorithms are rather slow but they ensure to output the perfect alignment for a given scoring matrix. Considering the sizes of some large protein sequence database, the usability of dynamic alignments is limited. Thus, alternative strategies were developed for the use of database search. In 1988 Pearson and Lipman introduced a heuristic approach for global alignment termed FASTA [57], which was supposed to speed up the alignment without sacrificing reliability. By focusing only on long identical segments between the two aligning sequences, the algorithm ignores the remaining alignment space. This results in a significant increase of speed due to the smaller alignment space with an acceptable risk of not finding the optimal alignment. Few years later, Altschul et al. introduced another heuristic approach for a local alignment termed BLAST [58]. BLAST (Basic Local Alignment Search Tool) was mainly developed for the use on large datasets. By dividing the query sequence into overlapping words (default are 3 amino acids for protein sequences and 6 bases for nucleotide sequences), BLAST generates a list of all consisting words and adds a list of similar words with a certain threshold for similarity. With these decoys the indexed database containing a large number of sequences is now queried. Only those sequence entries containing the sequence of the decoys will be considered for the alignments. Extending the alignment of the decoy hits in both sequence directions leads to high-scoring segment pairs (HSPs). These are stopped once the alignment score falls below a specified threshold. Figure 12 shows an overview of the algorithm. The speed increase with respect to dynamic alignment algorithms is highly significant due to the drastic reduction of alignments and the alignment space. BLAST also contains a module for statistical analysis estimating the significance of calculated similarity for a given alignment based on the Karlin-Altschul statistics [59]. This allows the rating of the similarity and may lead to homology estimations. Computationally, this is done with calculation of two values: the P- and E-value. The P-value indicates the probability that a given similarity score between two sequences occurs with the same or higher value also in other sequence alignments and thus does not have a high significance. Its corresponding E-value is the number of expected sequences with the same or even higher similarity in a database with a given number of sequences.

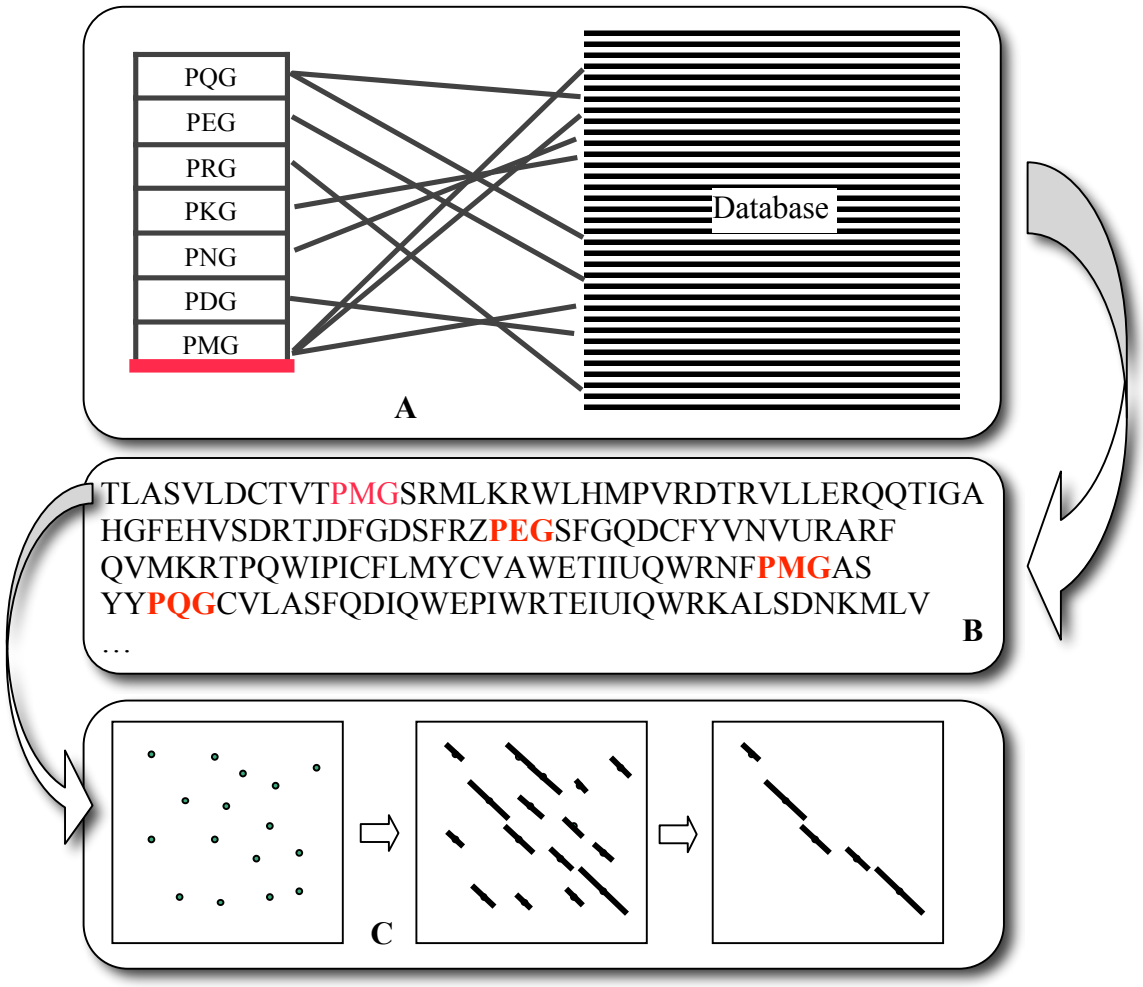


Figure 12: *BLASTP procedure. (A) The query sequence is divided into overlapping words in the size of 3 letters. By defining a similarity cutoff (red line) similar words will be added to this list. The database is now queried with the decoys of this list. (B) Sequences with matches are retrieved. (C) Each retrieved sequence is aligned to the query sequence where each word-match is extended until its score falls under a defined threshold. These HSPs will be reduced to the e.g. longest 10 and the overall alignment is given by the HSPs lying nearest to the diagonal of the alignment box.*

1.2.3.2. Multiple Sequence Alignments

Finding motif among functionally or structural related sequences has become an interesting research field especially in the area of protein classification and domain characterization based on the protein modularity assumption mentioned before. For this purpose multiple sequence alignment-methods were introduced in the early 1980s. Similar to the use of pairwise sequence alignments for database queries, dynamic

programming cannot easily be extended to more than three protein sequences as it may become enormously expensive in computing time. Therefore, heuristic methods were developed. A common method for performing a heuristic alignment search is the progressive technique. It constructs a multiple sequence alignment by first performing a series of pairwise alignments. The two most closely related sequences are first aligned and the next most closely related sequence is then successively aligned to the previous alignment as shown in figure 13. This also leads to a major limitation of progressive methods, which is their dependence on the initially assigned relations among the sequences and on the quality of the first alignment.

A popular progressive alignment method is the Clustal method, especially the weighted variant ClustalW [60] where the scoring function is modified with a weighting function that assigns scaling factors to individual members of the query set based on their relation distance from their nearest neighbors. This modification leads to a weaker effect of relatively poor initial alignments early in the progression. However, since progressive methods are heuristic methods that do not guarantee to find a global optimum, the alignment quality is difficult to be evaluated and biological significance may not be implied.

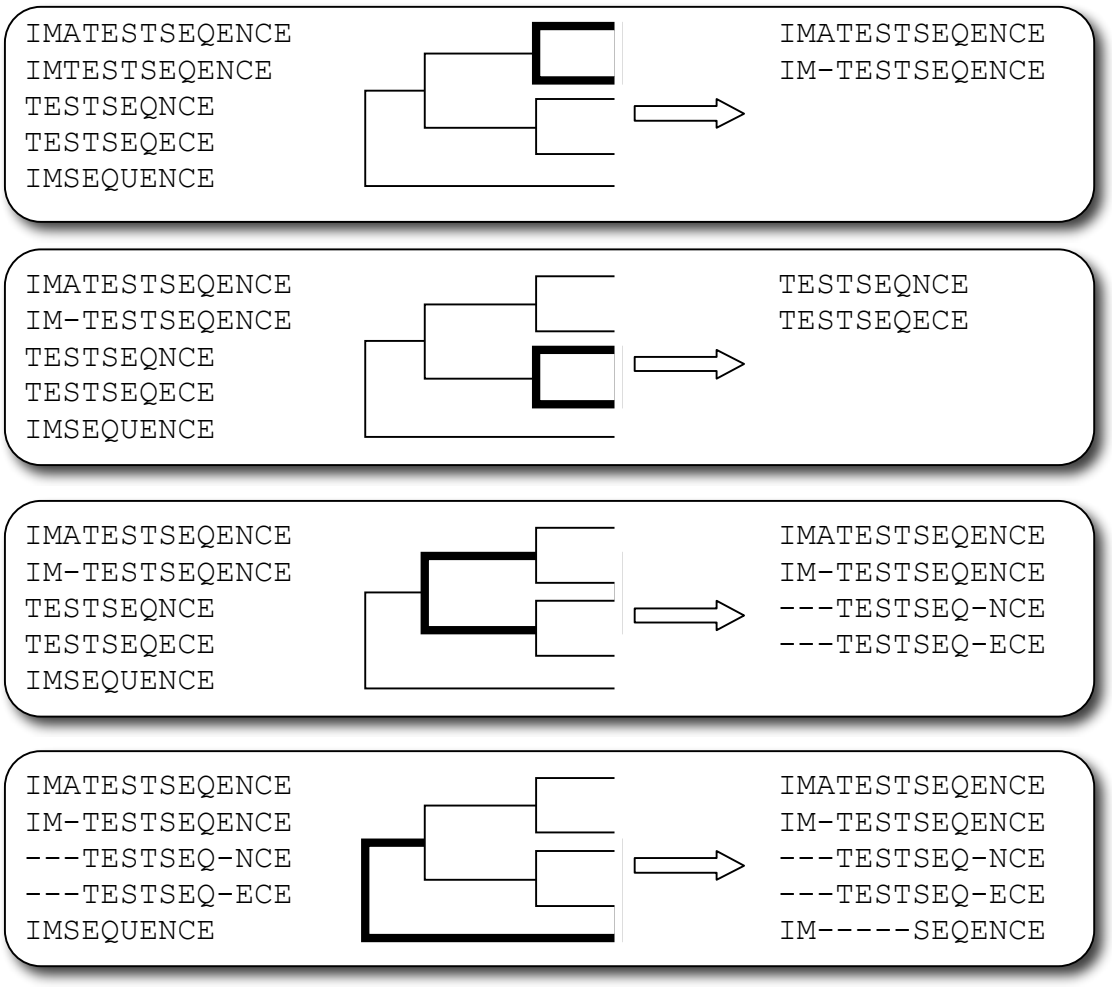


Figure 13: *Progressive Alignment. Each bin contains a part of the progressive alignment that is oriented in the order of similarity. The first two sequences are most similar and initiate the progressive alignment. Sequences 3 and 4 are most similar to each other and are aligned separately. In the 3rd alignment a consensus sequence of each previous alignment is used. Any evolutionary change that is applied to the representing sequence will be applied to its previous sequences as well. Finally, a representing sequence from these previously aligned sequences is aligned against the 5th sequence. This results in a multiple sequence alignment.*

1.2.3.3. Consurf [61]

Deriving conserved sequence regions was addressed in the previous section. Here, the well-known tool Consurf is presented. It calculates residue conservation scores and additionally projects these scores on a three-dimensional protein structure. The program package consists of several tools. In the first step Consurf collects a number of

homologous sequences to the query sequence using the non-heuristic Smith & Waterman local alignment algorithm. Given a threshold for the E-value, all sequences within this threshold are collected and duplicates discarded. In the next step a multiple sequence alignment is performed among these homologous sequences using ClustalW (figure 14). Based on specific rules for scoring amino acid-exchanges and gap penalties for insertions or deletions, the program calculates an average score for each position in the query sequence and applies normalization for each score. This is necessary since the scores provide a reference state for the level of conservation. The normalization is based on the authors' assumption that surface residues that are involved in interactions with other molecules should be as conserved as the internal residues determining the protein structure. Therefore, a residue that is detected by Consurf as the most conserved is considered as conserved as a residue that is buried in the core of the protein. Subsequently, the program replaces the temperature B factors in the input PDB file with the conservation grades of the residues, which allows the conservation-mapped protein structure to be viewed in most pdb-viewers (see 1.2.4.1).

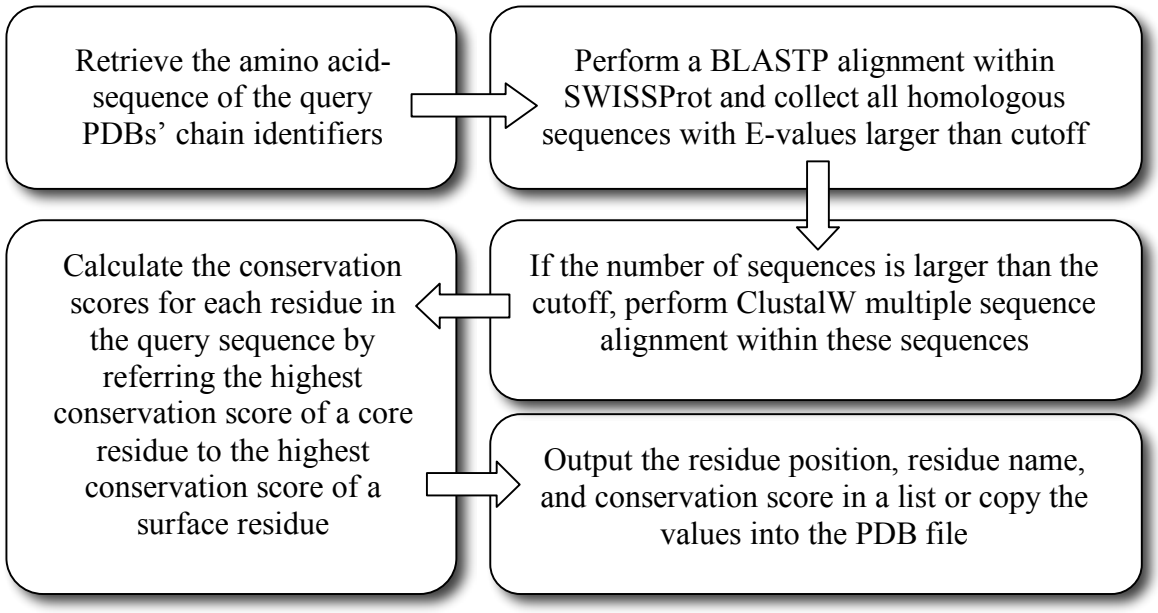


Figure 14: *Consurf procedure.*

1.2.4. Structural Analyses

Previously, some examples were described for analyzing the rich sequence data. In this section the most informative data is discussed: the protein structure data. Although the number of available structural data is quite limited, their information content is very high so that one can still obtain interesting insight into protein folding and protein binding. The connection between fold and function especially makes clear that understanding the function of proteins based on structural information should be the most straightforward way.

The fold of proteins is mostly defining their molecular activities. In fact, the fold reveals binding sites, interaction surfaces and the precise spatial relationships of catalytic residues. However, the connection between function analysis and structural data is not always clear. One protein-folding topology may support a variety of functions and, conversely, one function may be associated with several different folds.

Due to the lack of available structures, especially for complexes that can be used for analyzing interactions between proteins, many approaches have been developed to predict their interaction sites (see section 1.2.4.3.). These interactions are key to understand biological processes. Although there are also a number of strategies to predict interaction area from sequence data only [62][63][64], determining the structure of biomolecular interfaces is the best basis for a wider understanding of biological processes. The knowledge on structure also provides the possibility to modify their molecular interactions via structure-based drug design, site-directed mutagenesis and protein engineering. Therefore, docking and other structural prediction methods play an important role in structural bioinformatics.

This chapter introduces the program package VMD that contains a large number of modules and functions, and the technique of protein-protein docking. Additionally, a short overview over current interface retrieval strategies is given.

1.2.4.1. VMD [65]

Visualization methods are important for displaying molecular properties on molecules. These tools mainly allow rendering molecules according to numerical properties, e.g. the hydrophobicity and charges on the protein surface. Although there is a large number of available applications, only one that was employed for visualizing and analyzing all protein-protein interaction data is discussed in this chapter.

VMD is a molecular graphics program designed for the display and analysis of molecules and molecular assemblies, in particular biopolymers such as proteins and nucleic acids. Aside from a large number of visualization options, VMD also includes a number of plugins, functions, as well as a terminal interface using the Tcl embeddable parser to allow complex scripts with variable substitution, control loops, and function calls. A typical script in VMD has 4 stages: 1. Load molecules; 2. Select atoms by given criteria; 3. Perform measurements and calculations on the selections; 4. Output the results. Stages 2 and 3 are based on VMD functions 'atomselect' and 'measure'. 'atomselect' is a function for selecting atoms for a given argument. The variety of arguments allows detailed selections, e.g. atoms of a specified protomer in multi-chain complexes. What is more interesting, 'atomselect' also includes arguments for calculated distances between atoms. Given a distance cutoff of 5Å, a very short argument such as 'chain A and within 5 of chain B' results in all atoms of chain A that are within 5Å of any atom in chain B. In the chapter 2 this function is used to retrieve interface residues for a given distance criterion. Additional functions allow the retrieval of further information such as the description of the amino acid to which the atoms belong to, and the secondary structure element of its amino acid, and more. The 'measure' function supplies several algorithms for analyzing molecular structures. Accessing the list of atoms collected with the 'atomselect' function allows 'measure' to compute the solvent accessible surface area for the selection. Combining selection list and results from calculations such as the solvent accessible surface area, one may separate surface atoms and residues from those that lie in the core, or in the interface region for a given criterion.

1.2.4.2. Interface Definition

When analyzing protein-protein interactions, previous authors mostly focused on the interface region only. However, as the diversity of the interface definition in different studies shows, there is no clear definition that non-controversially specifies this region [66][67][68][69][70]. Most of the definitions can be divided into two major groups: distance based and solvent accessibility criteria. In the distance dependent criteria, the distances between Cβ [66] or any heavy atom [67] of interacting chains are measured and those that lie within a specified distance threshold are understood to take part to the interface. The common distance threshold in the literature is 5Å. It describes a good compromise for discriminating relevant from irrelevant interactions. Relevant interactions are van-der-Waals and polar interactions. Although electrostatic interactions play an important role in interaction specificity and encounter steering, it becomes very challenging to consider such interactions by employing a large distance cutoff. Within the large interaction range a large number of irrelevant interactions may hardly be discriminated from those specific electrostatic interactions. Irrelevant interactions are also those that do not occur under physiological conditions and can be referred to crystal packing.

In this work the method for determining amino acids that are involved in interface regions is mostly based on definitions applied via customized VMD scripts. By calculating all distances between all atoms in the complex, a distance cutoff can be used as a filter. A list of atom pairs with a shorter distance from each other than the cutoff value is generated. In the next step all those atom pairs from the same chain are discriminated. At this point, the list contains all atom pairs within a given distance cutoff and that belong to different chains. Another interesting approach was proposed by Jernigan et al. who used a criterion based on counting atomic contacts between opposite amino acids and defined interface atoms, such as those with more than a certain number of atomic contacts [71].

Solvent accessibility criteria are based on the loss of solvent accessible surface area (SASA) upon complex formation. For instance, Zhu et al. specified a residue as an interface residue if the loss of its SASA is greater than $1Å^2$ [68]. Bahadur et al. defined a loss of SASA greater than 1% as an interface residue [72]. The SASA area of these

studies was calculated by NACCESS [73]. The program uses the Lee & Richards method [74], whereby a probe of a given radius is rolled around the surface of the molecule, and the path traced out by its centre specifies the accessible surface. Typically, the probe has the same radius as water (1.4Å) and hence the surface described is often referred to as the SASA. The calculation makes successive thin slices through the 3D molecular volume to calculate the accessible surface of individual atoms. Another approach uses VMD together with the 'measure' function for atom surfaces. The function 'measure' calculates a ball with a given radius around any atom. Then it computes the non-overlapping area. This method is very fast and can be applied to molecules larger than 20,000 atoms, which is the limit for NACCESS. However, in this work it turned out that as VMDs' measure does not consider the protein fold, it confuses internal cavities with surface patches.

1.2.4.3. Protein-Protein Docking

Computational protein-protein docking is a technique for predicting how one protein will bind to another. Given two proteins of identified structure that are known to interact, docking methods may determine their natural complexed structure. The information of how and where two proteins bind allows a large number of further studies. Most of them are related to the field of drug design. Although RCSB PDB already contains a number of protein-protein complex structures, crystallization of protein-protein complexes remains to be a very challenging process due to rather weak affinities between the protomers.

Performing a protein-protein docking first requires the structures of the two proteins. Given the complexity of the structures considered at the atomic level, a simplified description of the structure is typically constructed. For example, the protein structure can be reduced to a series of cubic elements by discretizing the three-dimensional space using a grid (figure 15). Defining the grid size regulates the level of detail: the larger the grid spacing, the blurrier the representation of the molecule. Discretized structures on a grid allow fast surface matching when using methods such as Fast Fourier transform. It suffices to consider the relative movement of one protein with respect to the other one that is kept fixed at the center of the grid. When considering a translational (x, y, and z) scan only, the mobile molecule B moves through the grid representing the static molecule

A and a function describing shape complementarity f_C is computed for each relative orientation. Mathematically, the correlation function $f_C = f_A * f_B$ is given by:

$$f_C = \sum_{x=1}^{N}\sum_{y=1}^{N}\sum_{z=1}^{N} f_{A_{x,y,z}} * f_{B_{x+\alpha,y+\beta,z+\gamma}}$$

where N is the number of grid points along the cubic axes x, y, and z and α, β, and γ are the translational vectors of the mobile molecule B relative to the static molecule A. Since f_A and f_B are both discrete functions representing the discretized molecules A and B, it is possible to calculate f_C more quickly with the Fast Fourier Transform requiring only $\log_e(N^3)$ calculations instead of N^3. However, after each translational step molecule B can also be rotated around its axes x, y, and z. For this step Euler angles are used to minimize the computational efforts. Euler angles are a set of angles for given step sizes that lead to unique structural orientations in the three-dimensional space.

Additionally, there are a number of algorithms employing heuristic methods for scanning the docking possibilities. Such methods are mostly considered when the computational efforts are high as it is the case for flexible docking. When one of the two molecules is not treated as a rigid-body but flexibly, conformational changes within the molecule are considered as well. However, most protein-protein docking approaches are rigid-body dockings, while flexible docking is rather applied for protein-ligand docking, where the protein is mostly held rigid and the small ligand is treated flexible.

In 1992 Katchalski-Katzir et al. introduced a rigid-body docking using a Fourier transformation [75]. The authors developed a purely geometric docking approach considering flexibilities at the interfaces by allowing surface penetrations. In 2003 Huang et al. implemented this approach using the BALL library [76] and labeled it BDOCK. Due to the purely geometric validation of calculated complex formations, an additional scoring unit was added and later modified by Kunz and coworkers. By collecting e.g. the top 2000 ranked structures based on their interface complementarity, an additional program rescored these structures by evaluating their residue compositions at the interface region. Similar approaches were used in other groups as well [77][78][79]. A common scoring function for re-ranking docking outputs from purely geometric docking approaches is RPScore [79]. This pair potential function was derived from observed intramolecular pairings in a database of non-homologous protein domains, as well as

from observed intermolecular pairings across the interfaces in sets of non-homologous heterodimers and homodimers. The authors also applied fraction methods and achieved a significant improvement of the docking ranks when compared to the ranks after the shape complementarity docking. Fraction methods compute a potential based on the logarithmic rate of the counted and expected values. There are different concepts to define the expected value. One common way is based on the frequency of a residue to occur in the protein, which is based on the different frequencies of occurrence of different residues [80]. The mole-fraction method is proportional to the product of the fractional abundances of the residues in the pair. The contact-fraction method on the other hand is proportional to the propensities of the two residues to be paired with any residue.

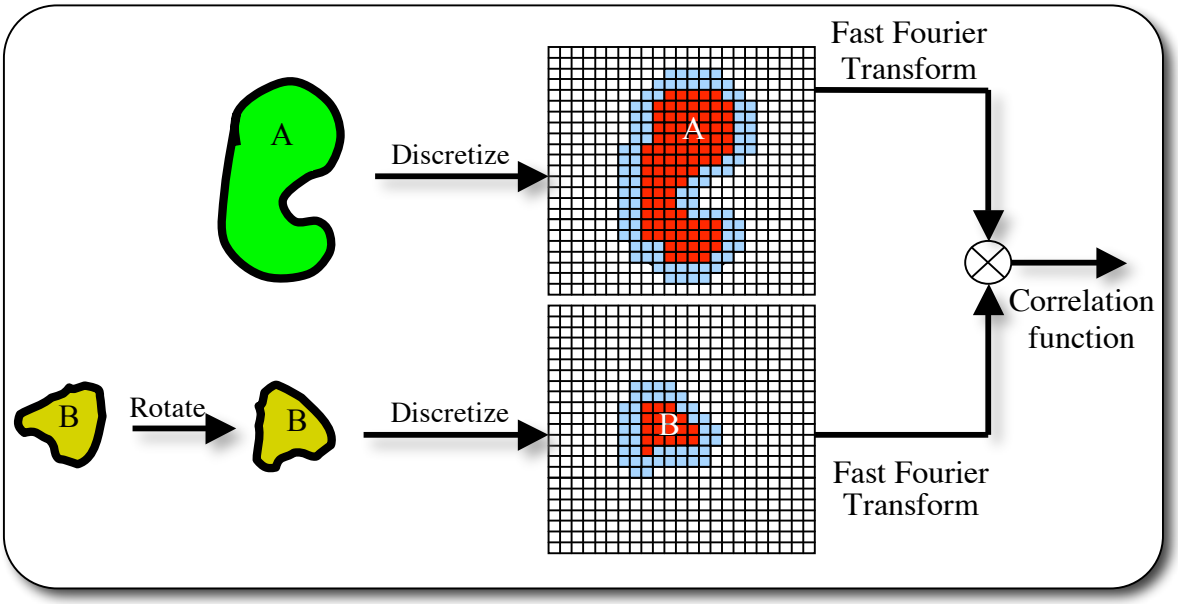


Figure 15: *BDOCK procedure. Given two proteins A and B, where B is the smaller protein, both protein structures are discretized into a three-dimensional grid (here only two-dimensional). Using the Fast Fourier Transformation all translational steps are applied to the mobile discretized protein B in the static grid of protein A in order to calculate the correlation of the two contact surfaces as it represents the geometric complementarity. This procedure is repeated for each rotational step over all three axes.*

1.2.5. Data Mining

Previously, a number of techniques were introduced that generated large amounts of data that could be exploited to gain a deeper knowledge on biological processes. Analyzing such large and diverse data requires the aid of computational methods. In this chapter two common data mining methods are described.

Data mining is defined as the process of discovering knowledge or patterns mostly hidden in large datasets. In the past years a large number of machine-readable datasets have literally led to a data explosion. Computational methods for extracting information from large quantities of different data are e.g. artificial neural networks, Bayesian networks, decision trees, genetic algorithms, statistical pattern recognition, support vector machines and others. Combining these methods with todays' larger computing power improved the analyses significantly.

Two main categories of data mining methods are common for analyzing protein function by a number of properties: clustering and classification. Clustering is used to organize a collection of unlabeled patterns into clusters of similar patterns. For a given model, these clusters will be most similar to each other than to other clusters. They yield clearer patterns from bulky data and ease their analysis. In the context of protein-protein interactions, clustering methods were used to identify clusters of e.g. different interface types for given interface properties.

As shown elsewhere, similar functions yield similar interface properties [17]. Therefore, the use of classification techniques may assign functions to interface properties. Instead of learning functional classification of proteins in an unsupervised way like clustering, classification techniques start with a number of pre-classified patterns. The goal in classifications is to find a model that will be able to categorize a new pattern.

1.2.5.1. Clustering

The goal of clustering is to group a given set of data points by their similarity. Next to the available data points, a system for estimating the similarity has to be employed. When it comes to constructing phylogenetic relations, sequence similarity or homology is mostly used for clustering. In the case of protein-protein interaction clustering, a similarity

measure such as Pearsons' correlation is used. Given the data and a measure for distances, a clustering method can then be used. There are two categories of clustering methods: hierarchical and non-hierarchical algorithms [81].

Hierarchical clustering is based on a hierarchy structure like a tree, which is basically an interlaced series of partitions e.g. in figure 13. The hierarchy is built from individual elements by progressively merging clusters. The first step determines which elements have to be merged in a cluster. Usually, the two closest elements are clustered first. Given the example in figure 13, sequence 1 and 2 are clustered first since their similarity score is highest. This results in the clusters: (1,2)(3)(4)(5). In the next step the distances between all elements are computed again. In the case of sequences 1,2 the average distance from 1 and 2 to the other elements (average linkage/UPGMA), the minimum distance from 1 and 2 to the other elements (single linkage/Minimum Evolution) or the maximum distance from 1 and 2 to the other elements (complete linkage) can be utilized. The output of such algorithms is an interlaced series of partitions that can be cut at any level forming a different partition. A popular example for a hierarchical algorithm is the Neighbor-Joining algorithm [82]. The principle of this method is to find pairs of close neighbors that lead at each stage of the clustering to a minimized total branch length. The algorithm therefore starts with a star-like tree.

Non-hierarchical clustering algorithms produce a single partition of the data instead of a clustering structure as a tree. Given large datasets, the complexity of such hierarchical trees can be high and inappropriate. K-means is the best-known partitioning algorithm. It starts with an initial partition and a fixed number of clusters and cluster centers and proceeds with assigning each element to its closest cluster center. New cluster centers are computed afterwards using the new cluster memberships. These steps are repeated until no changes are registered in the cluster memberships.

Since the validation of clustering results is difficult and the efficiency of a given clustering algorithm depends on the clustered data, there is no optimal strategy for clustering data points.

1.2.5.2. Classification

The idea behind constructing classification models for sample data is to train a system that can successfully classify new data. To estimate their predictability, the sample data is therefore randomly divided into a training set and a test set. Although classifications are mostly applied to large and labeled datasets, many recent studies also analyze rare structural data [83][84][85][86][87][88][89][90][45][46]. In that case, the division into training and test set is quite undesirable. A more suitable way to deal with this problem is the use of resampling techniques such as cross-validation. Taking the leave-one-out cross-validation, one data sample is taken out as a test sample while the remaining samples are used for the training. Systematically taking out each data sample for testing while training the remaining set will lead to an average prediction accuracy for a classification model. This way the full use of the limited number of data samples was assured for testing as well as for training. However, depending on the dataset this validation may become computationally very costly and inappropriate for some systems. In these cases the k-fold cross-validation is used where the dataset is randomly partitioned into k mutually exclusive test partitions and k-1 partitions are used for the training. The average error rates over all k partitions are then the cross-validation error rate. Mostly the 10-fold cross validation method is used.

There exists a large number of classification methods. Commonly used methods are Bayesian classifiers, linear discriminant analysis, nearest neighbor classification, classification tree, regression tree, neural networks, genetic algorithms, and very recently support vector machines.

Support vector machines classify data samples into two classes by fitting hyperplanes between the data points (figure 16). Although there are often many possible hyperplanes, the optimal hyperplane classifier stands in the focus of interest. The maximal margin of a separation can be uniquely constructed by solving a constrained quadratic optimization problem involving support vectors, a small subset of patterns that lie on the margin. The support vectors, often just a small percentage of the total number of training patterns, contain all relevant information about the classification problem. Figure 16 shows a simple partitioning of ‘O’ and ‘X‘ data. A linear separator can be constructed to separate the two classes as indicated by the red line. In cases where the SVM cannot linearly

separate the two data samples, non-linear decision rules (also called kernel functions) can be applied. Such functions map the data points into a high-dimensional feature space and then construct a linear separating hyperplane with maximum margin.

Support vector machines have become very popular in the area of classification since they always find a global minimum, while other strategies may get stuck in local minima. Its simple geometric interpretation is easily computed and can be applied for many cases. Similar to the clustering algorithms, no particular SVM kernel function is guaranteed to lead to sensitive classifications.

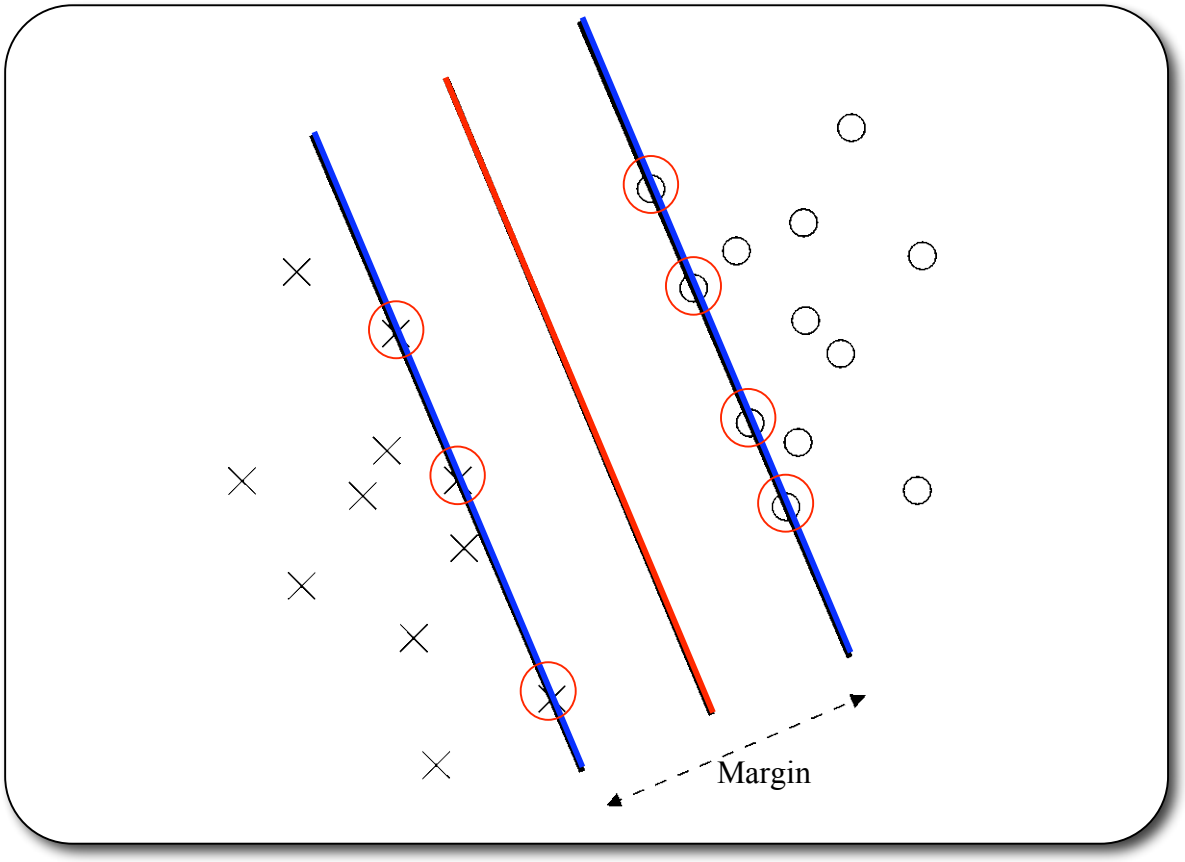


Figure 16: *Maximum-margin hyperplanes (blue lines) for a support vector machine trained with samples from two classes (circles and crosses). Samples along the hyperplanes marked with red circles are called the support vectors. The red line is the linear separator for this training sample.*

2 Statistical Analysis of Transient Protein-Protein Interfaces

2.1. Overview

In this chapter a non-redundant set of 170 protein-protein interfaces of known structure was collected and statistically analyzed for residue and secondary structure element-compositions and pairing propensities, as well as for side-chain and backbone interaction frequencies. A major goal of this work is to combine a number of previously analyzed aspects of protein-protein interactions to get a deeper insight in their nature. By now this was not possible, since most previous studies were based on different interface criteria, different and partially improper datasets, and different foci on types of interactions. The results of this chapter were published in the journal 'PROTEINS: Structure, Function, and Bioinformatics' in August 2005 [67].

2.1.1. Analysis of Protein Interfaces

Early statistical studies on protein-protein interactions have compared the compositions of internal and external interfaces [24][91][71][66][92][25][93]. Due to the small number of available structural data for protein-protein complexes, most of the studies did not distinguish between homo-multimers and hetero-multimers, as well as between permanent and transient interactions. This led to contradictory observations. Some studies showed a large dependency of residue composition on the type of the interfaces [91][25], whereas other reported that the residue compositions of different types of interfaces are rather similar [71][66][94]. Ofran and Rost introduced six types of interfaces [17]. For each interface type, they reported characteristic residue composition and residue pairing propensities.

2.1.2. Packing of Interfaces

A generally accepted conclusion from analyzing known structures of protein-protein complexes is that interacting proteins have a high degree of surface complementarity [95]. Tight packing of structural elements is therefore observed inside and between proteins [96][97]. It is interesting to look at the role of geometric complementarity in the packing of secondary structure elements as well. Jiang et al. characterized the role of geometric complementarity in secondary structure element-packing using a systematic docking procedure in order to recreate the crystallographically determined packing of secondary structure elements in known protein structures [98]. The apparent importance of the geometric match allowed prediction of the correct packing of the secondary structure elements based on a geometric fit alone. From high to low, the best packing were β-sheet and β-sheet, loop and loop, α-helix and α-helix, and α-helix and β-sheet. Such interface packing differs from core packing. Richards found that in known structures, core residues fill almost all the available interior space with minimal geometric strain and no steric overlaps [99]. Such dense packing is thought to provide many favourable van-der-Waals interactions as well as exclusion of solvent and thereby maximizing hydrophobic stabilization. Considering the comparably less hydrophobic surface and interface regions and the interfacial water molecules [25], the interface packing is probably not as tight as the core packing.

2.1.3. Transient Binding

The structural and thermodynamic basis for protein folding, protein assembly and protein-protein interactions are non-covalent contacts between residue side-chain and backbone atoms. Such contacts enable a large variety of associations and interactions within and between proteins. Since secondary structure elements α-helix and β-sheet are stabilized by hydrogen bonds between backbone atoms, these elements are obviously not residue-specific. In order to adopt their individual folds, protein structures are also stabilized by favourable backbone – side-chain and side-chain – side-chain contacts. Further examples for non-covalent interactions are such as those between side-chains of

separately folded chains, which lead to the assembly of multi-chain proteins. These are expressed in permanent interactions. In their review, Nooren and Thornton distinguished between transient and permanent complexes [16]. Their definition considers the association strength of complexes but not the environmental relation between the bound and unbound state. To incorporate this aspect, Nooren and Thornton used an additional classifier: obligate and non-obligate complexes. Here, the term "transient" is said to also fulfill the requirements for non-obligate and the term "permanent" for obligate interactions, since antigen-antibody interactions are discriminated.

2.1.4. Different Interface Sizes

Apparently, due to the lack of sufficient data only few attempts were made in the literature to distinguish the properties of rather small, middle, and large interfaces. In the process of analyzing the distributions of two groups of interfaces of different size, Glaser et al. found that hydrophobic residues occur more often on large contact surfaces, while polar residues prevail on small surfaces [66]. The exception is arginine, which is more common at large than at small contact surfaces.

2.2. Methods

2.2.1. Collecting Transient Protein-Protein Complexes

In order to collect structural information on transient interfaces, all multiple-chain protein entries in the PDB (September 2003) containing at least two chains each with a length of more than 10 residues were examined. Furthermore, discriminating criteria were employed for ignoring glycoproteins, carbohydrates, DNA/RNA and any DNA/RNA hybrids. Structures with resolutions lower than 3Å were skipped. To reduce the large number of non-complexes in the remaining list, the term "complex" was required to occur in the PDB header of the entries. Removing all homologous sequences at a level of identity higher than 90% (default setting on the RCSB PDB site) led to a set of 286 PDB files. Ensuring that the dataset included the desired complexes with the correct chain

identifications for the interaction, all structures were semi-manually examined and therefore approximately 130 complexes overruled. This step was necessary since a large number of permanent complexes, small ligand complexes, and antibody-antigen structures still passed the previous filters. The antibody-antigen interactions were not considered here because of their rather ambivalent classification, where the connection strength leads to rather permanent interactions although the protomers occur stably in their unbound state (non-obligate complexes). Another reason for not considering antibody-antigen interactions is based on the variable regions or complementarity determining regions containing highly variable residues that form loops. This will probably shift the statistical results for identified propensities in pairing of secondary structure elements. However, Lawrence and Colmans' study on the shape complementarity at protein-protein interfaces observed that antibody-antigen interfaces as a whole exhibit poorer shape complementarity than it is found in other systems involving protein-protein interactions [100]. This can be understood in terms of the fundamentally different evolutionary history of particular antibody-antigen associations compared to other systems considered in the study, and in terms of the differing chemical natures of the interfaces.

In order to enrich the dataset, another 59 complexes from the ZLAB benchmark set [101] for protein-protein docking were added. These complexes passed the examinations and resulted in 153 PDB files involving 170 interfaces and 24,290 residue pairs. The list of transient complexes is shown in table 4 and the composition of their functional classes is illustrated in figure 17. The dataset is obviously enriched by enzyme complexes, which includes the group of enzyme-inhibitor complexes. These interactions are very strong and may be defined as permanent interactions. However, as enzyme-inhibitor complexes are regulatory elements in biological systems, their dissociation is required. They are therefore counted as "transient" interactions.

1A3E-H L	1AY7-A B	1BRB-E I	1CXZ-A B	1FGL-A B	1HJA-I B	1LDT-T L	1SFI-A I	1UDI-E I	2R1R-C F
1A46-H L	1AZS-C B	1BRC-E I	1D4V-A B	1FIN-A B	1HJA-I C	1MAH-A F	1SGP-E I	1UEA-A B	2SEC-E I
1A4Y-A B	1AZZ-A C	1BRS-A D	1DAN-H T	1FLE-E I	1HKE-A D	1MCT-A I	1SIB-E I	1UGH-E I	2SIC-E I
1A5G-H L	1AZZ-A D	1BTH-H P	1DAN-L U	1FLT-V Y	1HLU-A P	1MEE-A I	1SLU-A B	1VAD-A B	2SNI-E I
1A81-A B	1B6C-A B	1BVK-A C	1DFJ-E I	1FMO-E I	1HWH-A B	1MKW-H K	1SMP-A I	1VIW-A B	2TEC-E I
1ABO-A C	1B7Y-A B	1BVK-B C	1DHK-A B	1FPC-H I	1IRA-X Y	1MTN-B D	1SPB-P S	1VRK-A B	2TGP-Z I
1ABR-A B	1BBZ-A B	1BVN-P T	1DN1-A B	1FQ1-A B	1ITB-A B	1MTN-C D	1STC-E I	1WQ1-R G	3EZE-A B
1ACB-E I	1BCK-A C	1CA0-B D	1E0A-A B	1FSS-A B	1JST-A B	1NOC-A B	1STF-E I	1XDT-T R	3HHR-A B
1AFE-H I	1BDJ-A B	1CA0-C D	1E96-A B	1GFW-A B	1JSU-A B	1NS3-A C	1TAB-E I	1YDR-E I	3HHR-A C
1AHW-A C	1BGX-T H	1CBW-B D	1E9H-A B	1GGR-A B	1JSU-A C	1NSG-A B	1TAW-A B	1ZBD-A B	3R1R-A D
1AHW-B C	1BGX-T L	1CBW-C D	1EAW-A B	1GL0-E I	1JSU-B C	1PDK-A B	1TBR-H R	2BTF-A P	3SGB-E I
1AK4-A D	1BI7-A B	1CDK-A I	1EAY-A C	1GLA-F G	1JXP-A C	1PYT-A C	1TCO-A C	2FAP-A B	3SIC-E I
1AN1-E I	1BI8-A B	1CEE-A B	1EBD-A C	1GOT-B G	1KIG-H I	1PYT-A D	1TCO-B C	2KAI-A I	3TEC-E I
1ATN-A D	1BJR-E I	1CGI-E I	1EBD-B C	1GPQ-A D	1KKL-A H	1PYT-B D	1TFX-A C	2KAI-B I	3TGI-E I
1AVG-H I	1BMM-H I	1CHO-E I	1EFU-A B	1GUA-A B	1KXQ-A H	1QBK-B C	1TGS-Z I	2PCC-A B	4SGB-E I
1AVW-A B	1BMQ-A B	1CM1-A B	1ETH-A B	1HE8-A B	1KXV-A C	1QMZ-A B	1TMQ-A B	2PCF-A B	5SIC-E I
1AVZ-B C	1BP3-A B	1CSE-E I	1FAP-A B	1HIA-A I	1L0Y-A B	1SBN-E I	1TPA-E I	2PTC-E I	7CEI-A B

Table 4: *List of 170 transient complexes. PDB ids and chain-identifiers are shown.*

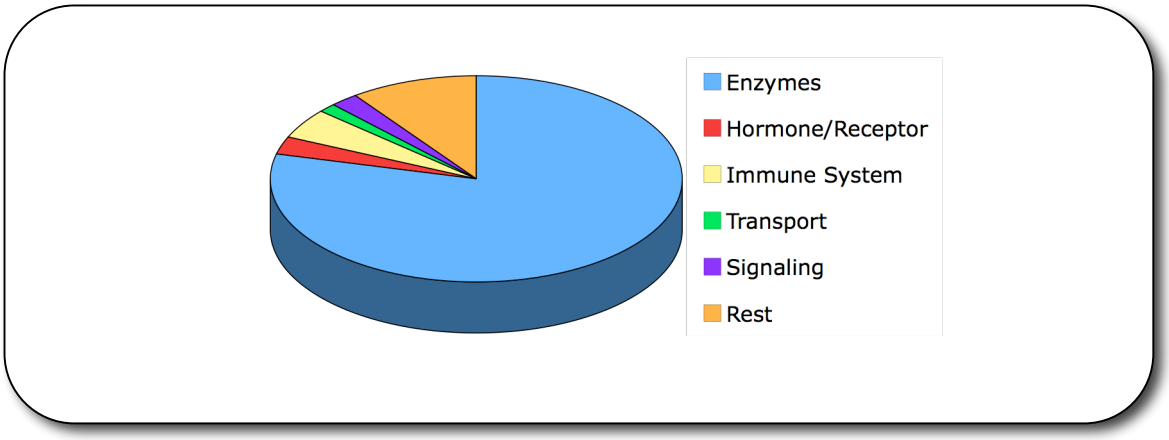


Figure 17: *Dataset composition of 170 transient complexes.*

2.2.2. Automated Analysis

Based on VMD each coordinate file was processed using a tcl/tk script retrieving all heavy atoms found within a given distance between two different chains. A residue is understood to form an interfacial contact in the case where the distance between any of its heavy atoms and any heavy atom from a partner chain is less than 5Å (also see section 1.2.4.1). This approach was also used by Aloy et al. [70].

Tracing the atom back to its corresponding residue allows analyzing interface residue-compositions. Given the amino acid-description of the atoms and the sequence position of the residues facilitated the retrieval of the corresponding secondary structure element

as assigned by VMD. This yields the secondary structure element-composition and pairing propensities. The interactions between side-chain and backbone atoms and their frequencies were analyzed as well. The distance criterion for this analysis was reduced to 3.5Å in order to discriminate non-specific interactions at the atomic level in a more satisfying way and yet keep sufficient data for significant statistical analyses. In this analysis the collected atom-pairs were traced back to the side-chain or backbone part of the corresponding amino acid. Finally, the dataset was split into differently sized interfaces. For that all interface participating residues were counted and their number was used for separating the interfaces into “small”, “middle”, and “large”.

2.2.3. Normalization

The statistics on pairing propensities of residues and secondary structure elements were normalized against the probability for a given residue or secondary structure element to occur at the interface given the following formula:

$$S_{ij} = \frac{c_{ij}}{\frac{n_i}{N} * \frac{n_j}{N}}$$

where S_{ij} is the score for the pairing propensity between the residues i and j or the secondary structure elements i and j. The value of c_{ij} is the number of binding pairs between i and j that occur at the interfaces of the dataset. The denominator is the product of relative frequencies of the residues or secondary structure elements i and j occurring at the interface. Disturbingly, the dataset turned out to be asymmetric. After investigating the original PDB files, it was found that some files (i.e. 1EAW) assigned several residues to the same sequence position, whereas the VMD program expects a unique residue for each position. To retrieve a symmetrical matrix without examining the whole content of the dataset, the arithmetic mean for both fields was then used differing by approximately 5% in the worst case.

2.3. Results and Discussion

2.3.1. Residue Composition at Interfaces

Two residues were considered to be in contact when the distance between any of their heavy atoms was less than or equal to 5Å. In this assay, all of the residues participating at the interface were counted. The average distribution of the entire dataset is 30.4% hydrophobic, 32.8% hydrophilic uncharged, and 36.8% charged residues. In opposite to other studies on protein-protein interfaces, the charged residues are the largest fraction [102][103][21][104]. Even the hydrophilic uncharged residues appear more frequently than hydrophobic residues. This finding attributes transient complexes that need to bind quickly and specifically but do not need to be stable for a long period of time and thus require a higher rate of hydrophobic residues. Comparing these results to large-scale studies [25][24][105][106][107][108][109] reveals the differences between dissimilar protein-protein interfaces. This is in agreement with Ofran and Rost [17]. It is not expected for permanent protein complexes to have a stable unbound state requiring a rather hydrophilic interface.

The tendency of some residues, such as methionine, tryptophan, and cysteine, to appear less frequent at protein interfaces agrees with the results and statistics of most other studies. Figure 18 shows a detailed graph for the interface distribution compared to the composition in SWISSProt. Methionine, tryptophan, cysteine, phenylalanine, tyrosine, arginine, and histidine are more strongly represented at interfaces. This finding generally agrees with those of Ofran and Rost focusing on hetero complexes [17]. Tyrosine and arginine are typically overrepresented in hot spots [110][24]. The enrichment of tyrosine as an aromatic residue can be explained by its ability to contribute binding energy through the hydrophobic effect without a large entropic penalty since tyrosine has few rotatable bonds. Furthermore, tyrosine is capable in forming multiple types of interactions in the lowered effective dielectric environment of hot spots, which is very favourable [110]. Besides tyrosine, a preference is also found for arginine, which may contribute to binding through electrostatic steering and is capable for multiple types of preferred interactions. Salt bridges can be formed with its positively charged guanidinium

motif, and the guanidinium π-system allows a delocalization of the electron, which leads to an aromatic character. It also has the ability to form hydrogen-bond networks with up to five H-bonds. The high preferences for arginine could also be explained with the ability of arginine to "guide away" water molecules from the interface during complex formation, or, conversely, upon dissociation. Pairs of aromatic amino acids tend to be preferred due to the π-π stacking. The higher occurrence of methionine, phenylalanine, tryptophan, cysteine, and histidine compared to the SWISSProt distribution could be a statistical balance of the under-representation of hydrophobic amino acids such as alanine, valine, leucine, and isoleucine. The under-representation of such hydrophobic amino acids at transient interfaces is very important. It ensures a stable unbound state and facilitates dissociations.

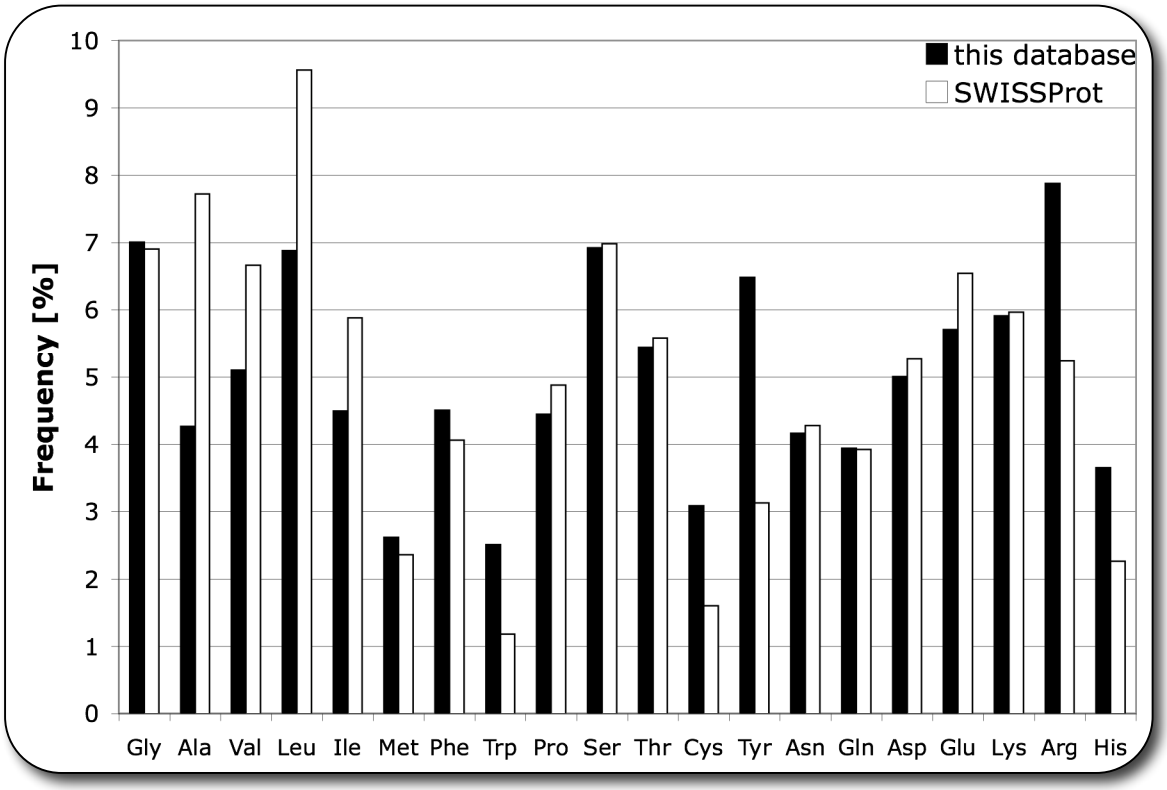


Figure 18: *Residue composition of protein-protein interfaces compared to the general composition in the SWISSProt database [111]. This data was retrieved from a distance criterion of 5Å between two interacting chains of 170 transient interfaces.*

2.3.2. Residue Pairing Propensity at Interfaces

Given the set of interface residues from the previous assay, each interface residue was re-selected and queried in order to find every corresponding residue in the binding chain that is located within 5Å to the selected residue. The computed statistics are given in a 20x20 matrix as shown in figure 19. These scores are normalized against the fractional abundance of each residue at the interface. For better visualization, a few representative rows for specific residues are shown more detailed in figure 20. Hydrophobic residues prefer to interact with other hydrophobic residues, which is evident from figure 20a. In contrast, pairs of hydrophobic and hydrophilic residues were associated less frequently compared to the number of hydrophobic–hydrophobic interactions, see figure 20b, while the charged residues showed very specific preferences according to their charge (figure 20c and figure 20d). Furthermore, the results support Glaser and coworkers' finding on very high association frequencies between tryptophan and proline as shown in figure 19. Such pairings are often found at the binding interfaces for proline-rich peptides on adapter domains like SH3. Another high score was observed for the interactions between phenylalanine and isoleucine. This is not surprising since both hydrophobic amino acids have rather flat and elliptic side-chains that have the ability to geometrically match.

As expected, one of the highest interaction peaks of figure 19 is found between arginine and glutamic acid. While the relative orientation of the charged groups of both residues suggests electrostatic attraction between both groups, a closer look reveals a broad range of residue-residue side-chain distances and angles reflecting a variety of electrostatic interactions, including salt bridges and hydrogen bonding. In addition to this, Glaser et al. also found that there is a hydrophobic interaction that may add to the pairing propensities of oppositely charged residues [66]. Even though these statistics show interesting but expected aspects of transient binding sites and underline the statistical strength of this study, the matrix does not correlate well with those of other studies, such as the RPScore matrix and Glaser's '*residue-residue contact preferences matrix*' [66]. Same happens between these two matrices. The highest peaks on phenylalanine and isoleucine are found in both matrices, as well as the favourable hydrophobic–hydrophobic and polar–polar interactions. The binding of proline and tryptophan is not as highly scored as in this

matrix, but the preference of charged residues, such as lysine and arginine for aspartic acid and glutamic acid, fits to these scores.

The dataset of Glaser et al. contains 621 interfaces, 404 of which are homodimers and 217 of which are heterodimers, including antigen-antibody interactions. The different character of the investigated interfaces may explain the low correlation with this study.

Ofran et al. showed that there are significant differences in residue composition and residue-pairing propensities between interactions of residues within the same structural domain and between different domains, between permanent and transient interfaces, and between interactions associating homo-oligomers and hetero-oligomers [17]. This leads to the assumption that the generalized data of RPScore and Glaser and coworkers' study may be the fundamental reason for the observed low correlation. This scoring matrix may therefore be more suitable for characterization of hetero-oligomer associations.

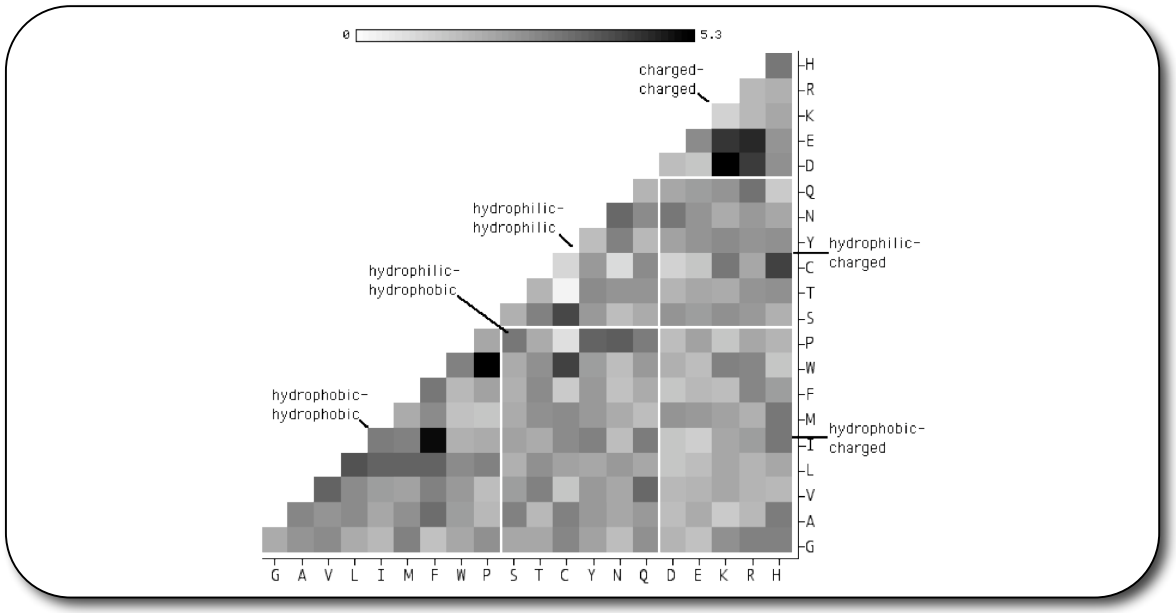


Figure 19: *Amino acid-pairing propensity matrix of transient protein-protein interfaces. Scores are normalized pairing frequencies of two residues that occur on the protein-protein interfaces of transient complexes.*

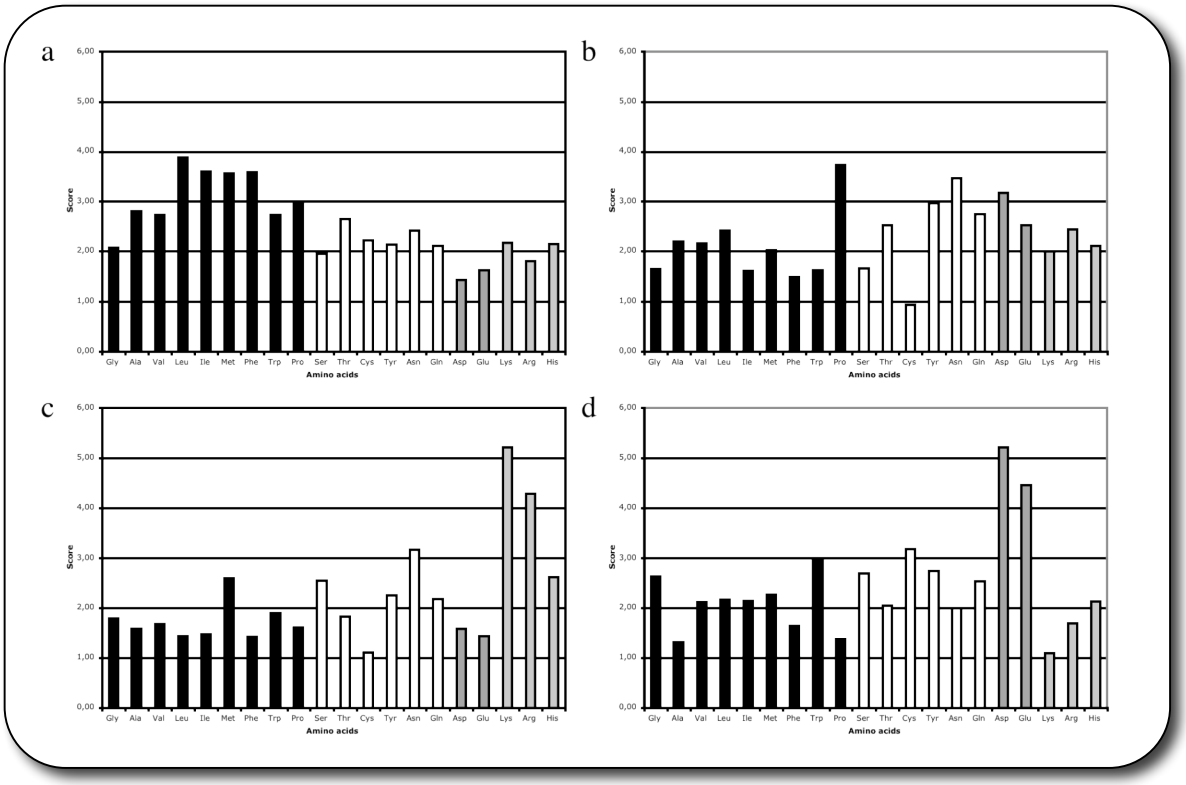


Figure 20: *Relative occurrence for binding partners of (a) leucine, (b) asparagine, (c) aspartate, and (d) lysine. Black bars indicate hydrophobic residues, empty bars hydrophilic residues, and grey bars charged residues. The higher the score, the more frequently such pairs occur in the dataset.*

2.3.3. Secondary Structure Element-Composition

The following analysis focuses on the types of secondary structure elements assigned to the interface residues. In addition to the interface composition, it might be interesting to study the types of secondary structure elements that are involved in these interfaces. The secondary structure element-composition is shown in figure 21. Helices and β-sheets occur infrequently, whereas turns/loops are overrepresented. Their statistical overrepresentation may be due to their ability to interact with different secondary structure elements. This suggests that interfaces need such bridges since larger secondary structure element-segments may come from the core of the protein and end at the surface or redirect into another secondary structure element-segment. The role of α-helices and β-sheets at the interface of transient interactions seems to be of less importance.

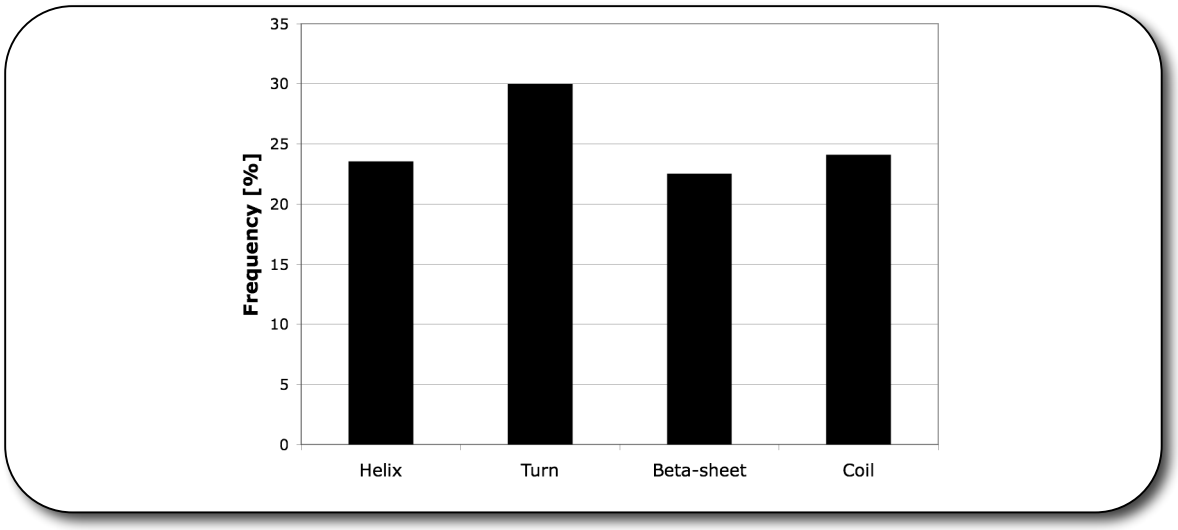


Figure 21: *Secondary structure element-composition at transient protein-protein interfaces. 'Helix' represents α-helix, 3-10 helix and π-helix.*

2.3.4. Secondary Structure Element-Pairing Propensity

The pairing propensities of secondary structure elements were collected from the residue propensity lists, simply by selecting the secondary structure element information for each of the residues. In the case of helices, including α-helices, 3-10 helices, and π-helices, and β-sheets the propensities are clear. As Jiang et al. ascertained, there is a strong preference for helix–helix and β-sheet–β-sheet interactions [98]. However, the results in figure 22 do not match in all cases. Whereas Jiang et al. reported that coil prefers coil the most, a larger preference between coil and turn/loop is found here. Interestingly, these results show a low pairing frequency between helix and β-sheet. Helix and β-sheet do not provide as tight packing as helix–helix and β-sheet–β-sheet do. This leads to the conclusion that the steric match plays an essential role in the packing of secondary structure elements, which is supported by other studies as well.

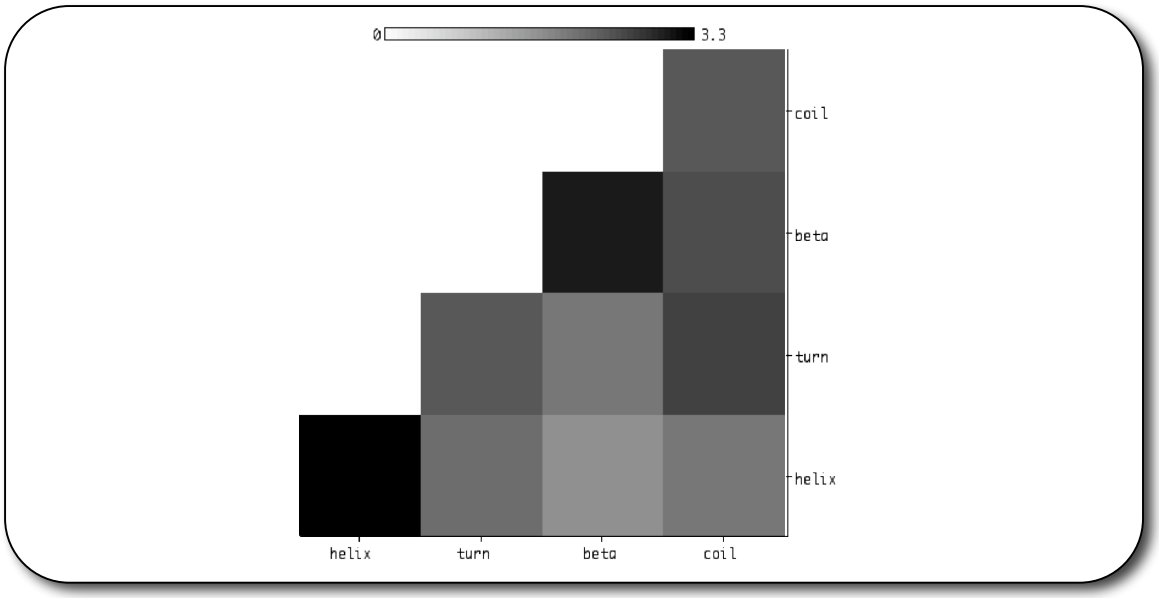


Figure 22: *Secondary structure element-pairing propensity matrix. 'Helix' represents α-helix, 3-10-helix and π-helix. Higher scores refer to higher pairing propensities.*

2.3.5. Side-Chain–Backbone Pairing Propensity

In order to enhance the precision and still retain statistically strong data, a tighter distance criterion of 3.5Å between the heavy atoms of each interface residue and the corresponding chain was chosen. This analysis shows that interactions between transiently bound proteins occur through a variety of backbone–side-chain contacts as reflected in figure 23. This agrees with findings of Aloy et al. and Jackson [70][112]. Additionally, the distribution of secondary structure element-pairings within certain binding-combinations is examined. Figure 21 showed that helices and β-sheets are not exceptionally overrepresented at interface area. To verify this, the next analysis focused on helices – including α-helices, 3-10 helices, and π-helices – and β-sheets and summed up all remaining secondary structure elements as 'else'. The previous findings are confirmed in this more stringent analysis, as shown in figure 23. Helix and β-sheet pairing combinations occur rarely, while the remaining, rather unstructured elements are more strongly involved in pairing combinations.

Figure 24 illustrates the preferred pairing combinations of secondary structure elements at a given side-chain and backbone interaction. Helix–helix pairs are more strongly represented

in side-chain–side-chain interactions. As already discussed, tight packing plays an important role in protein-protein interactions. Therefore, it is assumed that helix–helix pairs are preferred over helix–β-sheet pairs due to their ability to pack more tightly. Such helix pairs have more side-chain interactions involved than backbone atoms. On the other hand, the tightest packing for β-sheets is with β-sheets, which involves more backbone–backbone interactions. While helix and β-sheet pairs are not as tight in their packing, the interaction of side-chain and backbone atoms is quite balanced. With this study, the role of tight sterical packing at the interface region was underlined for transient protein-protein interactions, focusing on helices and β-sheets. In general, these two structural elements play a minor role when compared to the remaining, rather unstructured secondary structure elements. This underlines the concept that helices and β-sheets tend to emerge from the interior of the protein and are redirected to the interior by structures such as turns, loops or even coils. Long stretches of helical or β–sheet structures are highly unlikely to be part of an interface when it comes to transient protein-protein interactions.

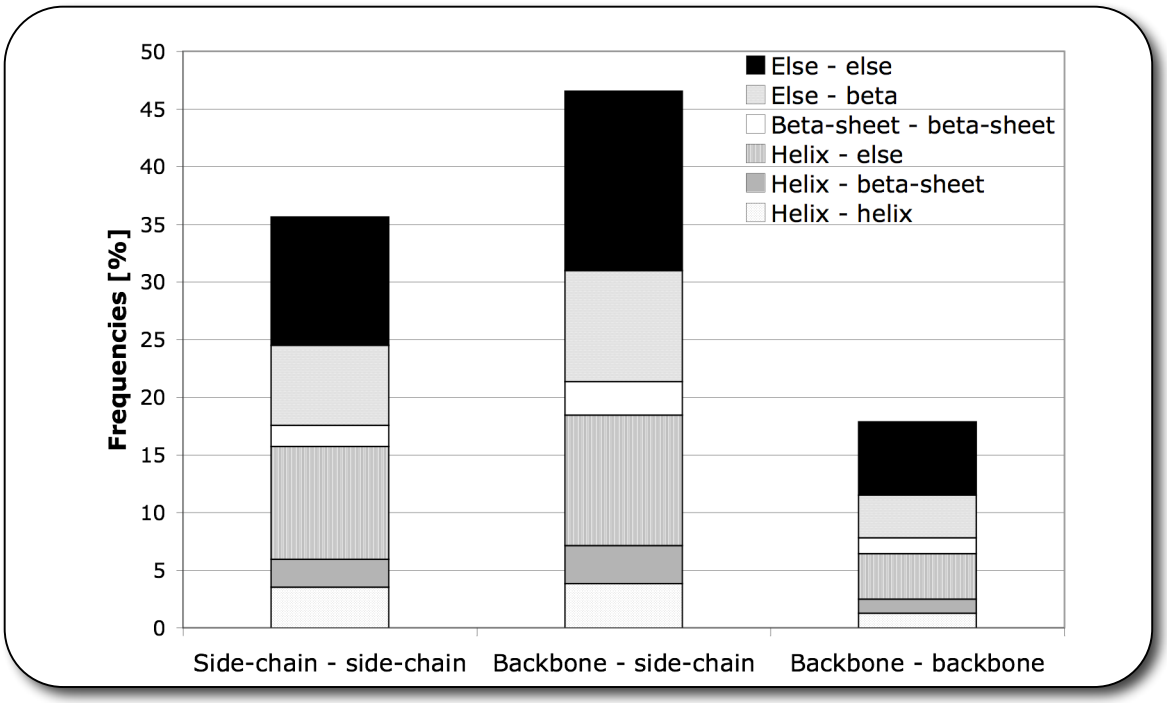


Figure 23: *Statistics on side-chain and backbone interactions and the secondary structure element-pairing propensities within given binding combinations. 'Helix' represents α-helix, 3-10-helix and π-helix while 'else' sums up any secondary structure element except helices or β-sheets.*

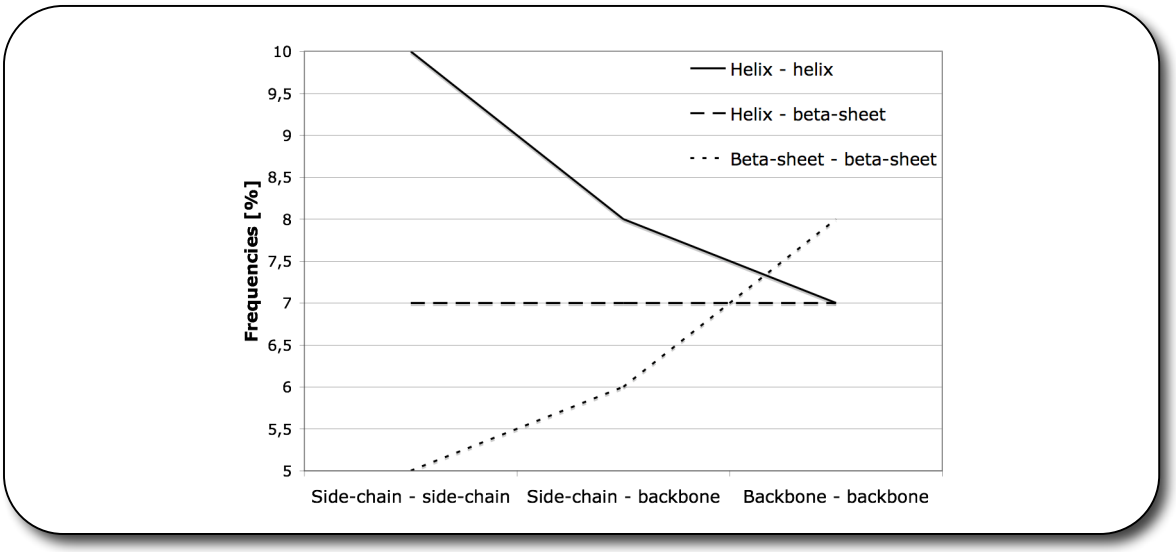


Figure 24: *Relative distribution of helix–helix, helix–β-sheet and β-sheet–β-sheet for all three binding combinations of side-chain atoms and backbone atoms. These frequencies derived from the distribution within one binding combination of figure 23.*

2.3.6. Comparison of Three Different Interface Sizes

This analysis is based on the interface residues initially counted at a distance criterion of less than 5Å. Figure 25 shows the number of interfaces of a given size, which is quantified with the total number of residues at the interface on both chains. As a rough separation, interfaces with less than 33 residues were defined to be "small", more than 32 but less than 68 residues to be a "medium"-sized, and everything beyond 67 residues was declared as a "large" interface. This separation was derived from the average and the standard deviation of this distribution.

Figure 26 shows the decreasing interface hydrophobicity as the interfaces become smaller. In general, hydrophobic residues contribute to binding affinity, but not as much to specificity. The opposite is valid for polar and charged residues. Small interfaces are characteristic for electron transfer complexes involved in energy metabolism where the two proteins need to bind quickly and with high specificity. Long lasting associations are neither required nor desired. Consequently, the frequency of hydrophobic residues is reduced and the number of polar and charged residues increased. On the other hand, large

interfaces rather need to be stabilized than specifically bound, which leads to the higher abundance of hydrophobic residues.

Interestingly, the curve of charged residues slightly drops when it comes to smaller interfaces. This may be an effect caused by the way interface residues are selected (here with a cutoff value for the distance between the atoms). Salt bridges between two charged residues may have distances larger than 5.5Å [70][113], which is beyond the threshold used here. The statistics could be "confused" at this point, considering that fast-associating and short-lived complexes prefer small interfaces and larger interfaces belong to slow-binding processes. The interpretation of these findings is hampered by the fact that kinetic and thermodynamic data is missing for many protein-protein interactions, or at least are not available in a convenient form.

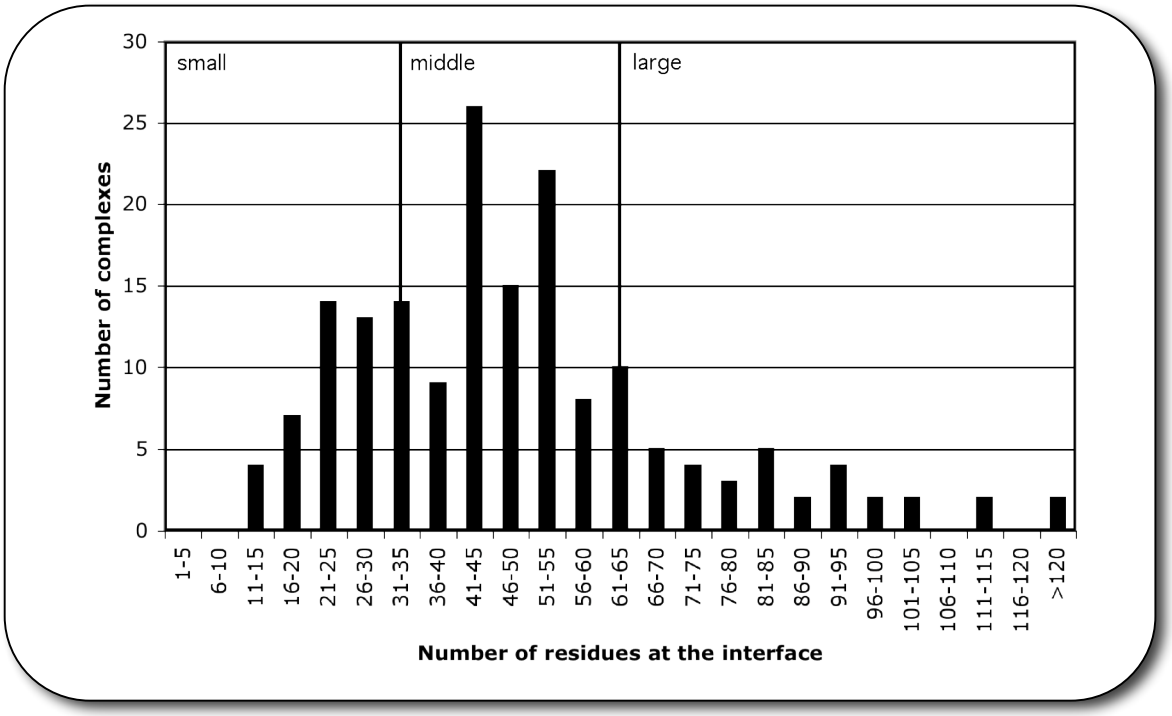


Figure 25: *Interface size-classification. This plot shows the number of complexes of a given interface size (quantified with the number of residues that are involved in the interface). The classification was derived by the average and the standard deviation of this distribution.*

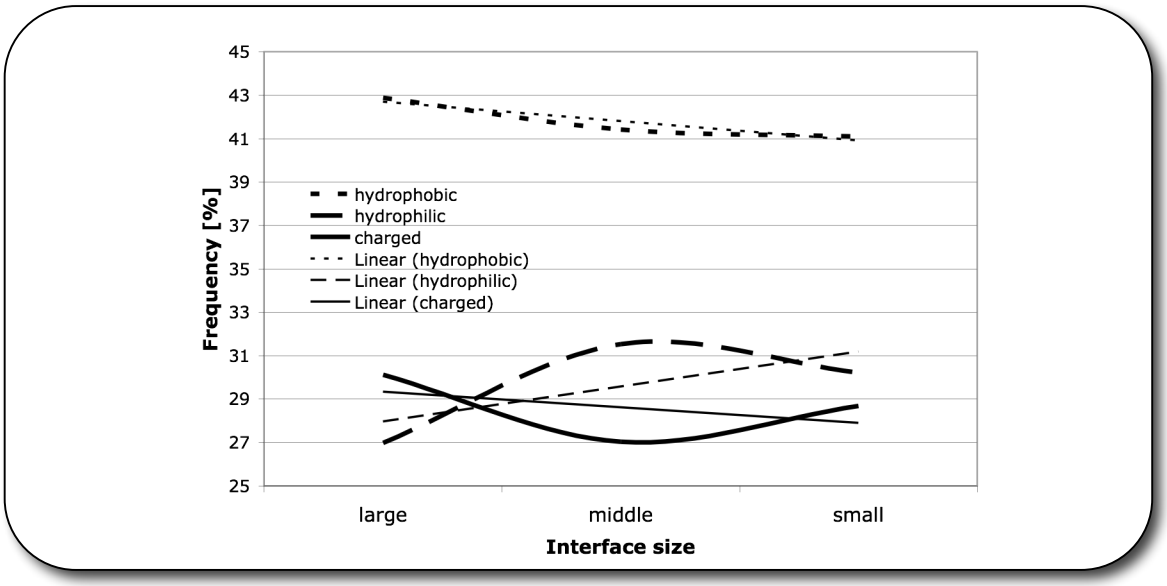


Figure 26: *Interdependency between interface quality and interface size. In addition to each interface-quality flow a linear trend graph is also shown.*

2.4. Conclusion and Outlook

Statistical information was collected on different properties of transiently bound interfaces. In general, the findings agree with those of previous studies as well as with the interpretation of experimental crystal structures. The differences from other studies found for interface-pairing propensities of residues most likely result from the focus of this work on transient interactions only. A new observation for the residue compositions is that charged residues dominate the distribution, and hydrophilic uncharged residues appear more frequently than hydrophobic amino acids. This emphasizes the importance of interface recognition rather than the stability of the complex guaranteed by hydrophobic residues. The results on pairing propensities of residues are as expected. Hydrophobic residues prefer interactions with other hydrophobic residues, while the charged residues show very specific preferences according to their charge. The analysis of the secondary structure element-content reveals that helices and β-sheets play a minor role at transient protein-protein interfaces. It is suggested that longer secondary structure elements come from the core of the protein and will be redirected at the interface leading to an enrichment of the rather unstructured secondary structure elements, mainly turns, loops and coils. A closer look at the secondary structure element-

pairing propensities shows the importance of packing, which is supported by other studies as well. Further analyses on side-chain and backbone interactions suggest high preferences between side-chain–backbone binding-combinations and underline the results on tight packing. In the studies on differently sized interfaces it was found that the hydrophobicity of interfaces drops as the interface becomes smaller. Generally speaking, hydrophobicity contributes to binding affinity rather than specificity. The opposite is true for polar and charged residues.

Finally, it should be noticed that the source of matrices for the pairing propensities of residues and secondary structure elements and additional analyses are based on the same method and dataset. This ensures compatibility between the different criteria and allows combination of the matrices as different steps in a filtering procedure. It is nearly impossible to derive comparable results from other independent studies, which are mostly based on different data and methods and hardly allow an overall conclusion. The next step will be to implement this information about compositions and pairing propensities of residues and secondary structure elements into a docking tool to test its performance of scoring protein-protein docking solutions.

3 Enhanced Sensitivity of a Docking Approach

3.1. Overview

In chapter 2 a non-redundant set of 170 protein-protein interfaces from transient complexes was collected and analyzed. A large number of characteristic properties were observed that may allow distinguishing between true transient complexes and other complex types or even crystal packing that form unspecific interactions. In this chapter a number of previously gathered information is used to test their predictability in finding native complex-formations proposed by a rigid-body docking approach. FFT docking was previously implemented in the Lenhof and Helms research groups by Hongbo Zhu and Bingding Huang using the BALL library. This implementation was termed BDOCK. Here, a modified version is employed and combined with residue and secondary structure element-pairing propensities in order to re-score highly complementary complex formations proposed by BDOCK.

The modification of the scoring function in BDOCK and the implementation of secondary structure element-scoring are based on a supervised FoPra thesis of Kerstin Kunz. This chapter will therefore focus on the benchmarks and analyses of the system.

3.1.1. The Rigid-Body Docking Problem

Starting off with the known three-dimensional structures of two proteins, protein-protein docking programs attempt to predict the three-dimensional structure of their complex. This became an important area in structural bioinformatics, as the number of experimentally determined protein structures rapidly increases and their complex formation often remains unknown. Keeping the unbound proteins rigid significantly reduces the computational time required for finding the optimal orientation of the two proteins. Katchalski Katzir proposed one of the most popular rigid-body docking

approaches in 1992 [75]. Discretizing the proteins on a cubic grid with given grid spacing and transforming the calculation into the Fourier space dramatically lowered the computational complexity for solving the search problem (see also section 1.2.4.3). It was previously stated that interacting proteins have a high degree of surface complementarity [95]. Tight packing of structural elements is therefore also observed between proteins [96][97]. This reduces the assessment of the orientations between the two proteins to the assessment of shape complementarity, which also reduces the computational complexity of the docking approach. However, given that proteins may undergo conformational changes once they form complexes and the evaluation of docking samples is now done just by shape complementarity, the algorithm is expected to produce a large number of false positive docking samples. There were a number of attempts trying to enhance the sensitivity of such type of docking approaches [114][115][116][117][118]. However, most of these solutions are time-consuming and therefore less appropriate for screening a large number of proposed complex formations.
In this chapter, an alternative approach will be tested that re-scores the proposed complex structures using given pair potentials for residues and secondary structure elements.

3.2. Methods

3.2.1. BDOCK

A weakness of rigid-body docking is the shape treatment based on rigid protomer structures, which may undergo conformational changes upon complex formation (induced fit). An obvious improvement of the rigid-body docking process therefore is the consideration of protein flexibility. In the case of keeping the protomer structures rigid, one should at least allow for some intermolecular penetration to mimic the effects of flexibility. However, such kind of flexibility consideration may only cover flexibility of side-chains but not of the backbone. The motions that constitute backbone flexibility are defined as hinge-bending [119] and were addressed for protein-ligand docking approaches [120] and just recently for protein-protein docking as well [121]. Another weakness may arise once complex formations are rated purely by their geometric complementarity. Given these

frailties, Huang and Zhu implemented the FFT docking algorithm in 2003 (Master thesis) using the BALL library and also introduced a subsequent screening function based on residue-residue pairing propensity-scoring in order to overrule most false positive structures.

BDOCK consists of a number of specifications and parameters that either lead to a detailed but slow, or to a blurry but fast prediction of complex formation such as the grid size, angle steps, and surface thickness for the surface penetration. Additionally, the native complex structure can be specified and used as an assessment for the predicted complexes computing their RMSD. The program outputs a specified number of predicted complex formations ranked by the order of their surface correlation score (see also section 1.2.4.3). The complex formations are described as the translational and rotational translocations of the mobile protomer in relation to the static protomer.

In 2005 Kunz and coworkers modified BDOCK and extended the scoring unit with secondary structure element-based scoring (FoPra thesis).

3.2.2. Docking Scoring-Function

Rigid-body docking approaches based on pure geometric complementarity do not guarantee to find the *in vivo* complex formation since their native formation does not necessarily rely on the geometric fit at the interface alone but also on biochemical complementarity. Thus, the developers of BDOCK extended the FFT docking implementation with a scoring unit that re-assesses the best complex formations based on their geometric complementarity by considering amino acid-pairing propensities. Although there are similar attempts based on atomic interactions [114][122], the use of residue-level potentials provides smoothness in the energy landscape that is likely to reduce the sensitivity of the function to precise atomic position. Additionally, these residue-based potentials are faster to evaluate.

Here, the scoring unit accesses a matrix containing residue-residue pair potentials. These potentials are based on counted residue pairs within the interface region for a given interface criteria and are statistically evaluated.

3.2.2.1. RPScore [79]

RPScore (Residue Level Pair Potential Score) is based on empirical pair potentials between amino acids. Each potential arises from pairing propensities of residues derived from interface-residue pairs within a given distance cutoff. Moont and coworkers collected a number of 103 non-homologous interfaces from the SCOP database. The authors specified, among others, three interface criteria and two fraction methods. A residue pair is selected if a specified distance cutoff between the atoms of interacting protomers is not exceeded for: (a) Cβ atoms, (b) any atom, or (c) the side-chain atoms. Furthermore, two different fraction methods were employed in order to retrieve residue-pair potentials. The potential calculation is based on a logarithmic ratio of the counted and expected pair:

$$S_{i,j} = \log\left(\frac{c_{i,j}}{e_{i,j}}\right)$$

where $S_{i,j}$ is the potential for the residue pair i and j. The value $c_{i,j}$ is the number of counted residue pairs i and j and $e_{i,j}$ marks the expected number of pairs i and j that can be calculated either with the mole-fraction or contact-fraction method. The mole-fraction method is proportional to the product of the fractional abundances of the residue in the pair:

$$e_{(mole-fraction)i,j} = C \times \frac{n_i}{N} \times \frac{n_j}{N}$$

where C is the sum of all obtained contact pairs and n_i/N, and n_j/N, the fractional abundances for i and j. On the other hand, the contact-fraction method is proportional to the product of the fractional contact propensity of the residue in the pair:

$$e_{(contact-fraction)i,j} = C \times \frac{c_i}{C_N} \times \frac{c_j}{C_N}$$

where c_i/C_N and c_j/C_N are the frequencies for i and j to be involved in any residue pair. The value of the score $S_{i,j}$ for each pair can be considered simply as a statistical measure of likelihood of that pair occurring. Since the quantity is a log fraction, the total likelihood for a structure is the sum of all the individual scores.

3.2.2.2. SARScore

Similar to RPScore, a residue-pair potential was calculated from the dataset of 170 transient interfaces. The residue-pairing propensities showed clear patterns, as well as the secondary structure element-pairing propensities and may prove useful discriminating false positive docking samples. Considering the rich amount of transient complexes and the larger dataset, the derived pair potentials can be expected to be more successful for scoring results from protein-protein docking than RPScore. In particular, this may be the case for unbound-unbound docking. Such protomers were structurally determined since their unbound state is stable. Complexes between stable protomers meet the criterion of non-obligate complexes, which is considered in the extended definition for "transient" as mentioned in chapter 2.1.3.

Pair potentials were derived for data obtained with four different distance cutoffs in order to specify interacting residues: 4Å, 5Å, 6Å, and 7Å. Similar to RPScore, the mole-fraction method and contact-fraction methods were applied to compute potentials. In the case of the mole-fraction method, the fractional abundances for a given amino acid were not calculated from the available data but retrieved from the SWISSProt statistics as shown in figure 18 [111]. It is assumed that the expected values become more accurate and the potentials more significant. For the contact-fraction, the fractional contact propensities of the residues collected from the available data were employed.

In a preliminary work of Kunz, the suitability of these two fraction methods was analyzed. Kunz found higher predictabilities of mole-fractioned residue-pair potentials based on FFT docking approaches. Based on this observation, residue-pair potentials are considered in the mole-fraction method only. However, Kunz also implemented the compatibility to a scoring matrix based on secondary structure elements. Similar to the residue-pair potential, the secondary structure elements of the residue-pairs within a given distance cutoff were collected and converted into pair potentials applying the contact-fraction method only. The mole-fraction method was skipped due to the missing data from larger datasets containing fractional abundances on secondary structure elements. Also, the results in figure 21, where the distribution of the abundances for given secondary structure elements is illustrated, show a nearly flat distribution which may not lead to useful mole-fractioned potentials.

Ultimately, the dataset of 170 transient interfaces was divided into ‘small’, ‘middle’, and ‘large’ sized interfaces, as described in chapter 2.3.6 and shown in figure 25. For each size, different pair potentials were computed based on the idea that the separation into differently sized interfaces may improve the docking sensitivity for matching docking samples.

Further in this chapter, the label SARScore (Structure And Residue Score) will be referred to as the mole-fractioned residue-pair potentials (SARScore(res)) and contact-fractioned secondary structure element-pair potentials (SARScore(struc)) based on the dataset of 170 transient interfaces.

3.2.2.3. Implementation of the Pair Potentials in BDOCK

Huang and coworkers implemented the FFT docking program using the BALL library. Using the BALL library facilitates simple implementation for the subsequent screening of docking samples by pair potentials. Specifying a cutoff for the best docking formations after their correlation value (default: best 2000), the complex formation is drawn from the given translational and rotational translocations of the mobile protomer and distances within all atoms are computed. If a computed distance between a given pair of atoms is below a cutoff value (default: 5Å), the corresponding amino acids are selected and the pair-potential value in the scoring matrix is retrieved. Summing up all pair potentials, a score for a given complex formation is calculated and used for re-ranking the complex samples from the docking as shown in figure 27. Kunz and coworkers programmed an additionally modified version of the scoring function that does not retrieve the amino acid-type for a given atom pair but its secondary structure elements stored in the PDB file. However, this may lead to incompatibilities since the secondary structure elements retrieved form the pair potentials are based on the module STRIDE [123] implemented in VMD and not on the secondary structure element assignments stored in the PDB file. Consequently, the PDB files of the protomers in the benchmark set were edited to store secondary structure element-labels computed by STRIDE.

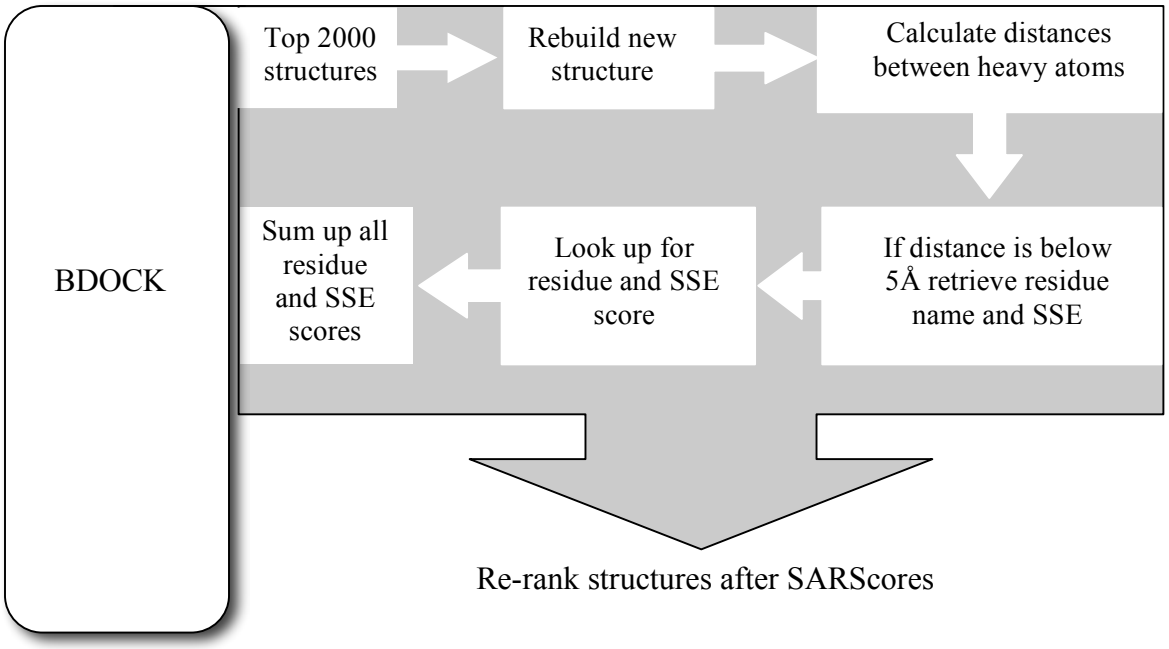


Figure 27: *Procedure of the scoring function for BDOCK. SSE stands for secondary structure elements.*

3.2.3. Benchmark

Benchmarking docking approaches can be divided into two classes depending on the input of protomer structures. If the known crystallographic complex structure is separated in two protomers and then docked into a complex again, it is called bound-bound docking. In this case, the protomers already have the complex fold and only need to be arranged correctly. In the case of unbound-unbound docking, the separately crystallized protomer structures are used for docking. In practice, the protomer structures may undergo conformational changes upon complex formation. A simple arrangement of the protomers may not be precise enough. These cases are not only more challenging for most rigid-body docking approaches, but also the typical application for docking.

In order to test the performance of the docking and scoring approaches, a set of protomers with known complex structures was taken from the ZLAB benchmark set 2.0 [124]. Although a number of complexes from the ZLAB were previously used to retrieve 170 transient interfaces, the newer version contains a number of new structures that are not included in the current database. This was the case for 8 complexes: 1EWY, 2MTA,

1F34, 1E6E, 1PPE, and 1D6R. Additionally, 9 benchmark complexes included in the dataset of 170 transient interfaces were added. This way the predictability of known interfaces was tested and compared to the 8 unknown cases. All together, 17 structures were used to benchmark the docking and the scoring. 15 of these structures are rather simple tests, where the unbound-unbound structures barely change conformations (RMSD bound vs. unbound <1.5Å). Two structures undergo larger conformational changes and mark difficult tests as shown in table 5. All structures belong to the class of enzyme-inhibitor complexes and should yield greater efficiencies when compared to the RPScore potentials as the dataset of 170 protein-protein complexes also contains a large number of enzyme-inhibitor complexes (figure 17). In order to estimate the efficiency of the docking and scoring approach, the docked complex structures are compared to the native complex structure. RMSDs below 3Å are defined to be near-native structures, as was also done by Huang and Schröder [125].

The basic parameters for these examinations are a grid spacing of 1Å, surface thickness of 2Å to consider flexibilities at the binding area, core overlap-penalty of -15, and angle steps of 10° for the rotations of the mobile protomer leading to 14,868 rotations over all three axis. As figure 27 shows, only the best 2000 structures ranked after their surface correlations will be considered for the scoring. Previously it was observed that most structures with low RMSD values to the native complex were listed within the top 2000 ranks.

3.3. Results and Discussion

3.3.1. Unbound-Unbound vs. Bound-Bound Docking

Bound-bound docking is based on protomer structures that are already in the conformation observed in the complex structure, while protomers with different conformations than in their complexed structure are used in the unbound docking. As conformational changes upon complex formation happen in most proteins complexes, unbound-unbound docking is applied in most of the cases. However, as rigid-body docking approaches barely consider conformational changes of the protomers, their usability for unbound-unbound docking is questionable. This aspect is analyzed here.

Focusing on 5 benchmark samples, the sensitivity of BDOCK for the given samples was benchmarked as unbound-unbound and bound-bound docking. In this section the best-ranked 500 samples were analyzed. Additional information on their degree of conformational changes, as shown in table 5, could reveal some interesting aspects. It is expected that the sensitivity of BDOCK is higher for those benchmark samples with rather low conformational changes of the protomers when compared to their complexed state. Figure 28 shows that in all 5 cases BDOCK produces more near-native complex structures when the protomers already have the bound conformation (white bars). RMSD values beyond 3Å mostly lead to similar distributions for the unbound-unbound and bound-bound docking. However, the figures in figure 28 show a slight trend, where – in the case of unbound-unbound docking (black bars) – easier benchmark samples with lower conformational changes (figure 28A) lead to more low-RMSD docking results than those with higher conformational changes (figure 28E). This trend is only disturbed with the results in figure 28C, which are unexpectedly bad producing mainly high-RMSD docking samples. The possible explanation for this outlier might be based on the rather large 10° angle step size, where important orientations could be "jumped over".

Benchmark sample	Conformational changes of the protomers upon complex formation [Å]
2SNI*	0.35
2SIC*	0.36
2PCC*	0.39
2MTA	0.41
1PPE	0.44
1AY7*	0.54
1EAW*	0.54
1MAH*	0.61
1EWY	0.8
1UDI*	0.9
1F34	0.93
1DFJ*	1.02
1D6R	1.14
1E6E	1.33
1HIA*	1.4
1CGI*	2.02
1ACB*	2.26

Table 5: *Conformational changes of the benchmark samples. *Data included in the dataset of 170 transient complexes.*

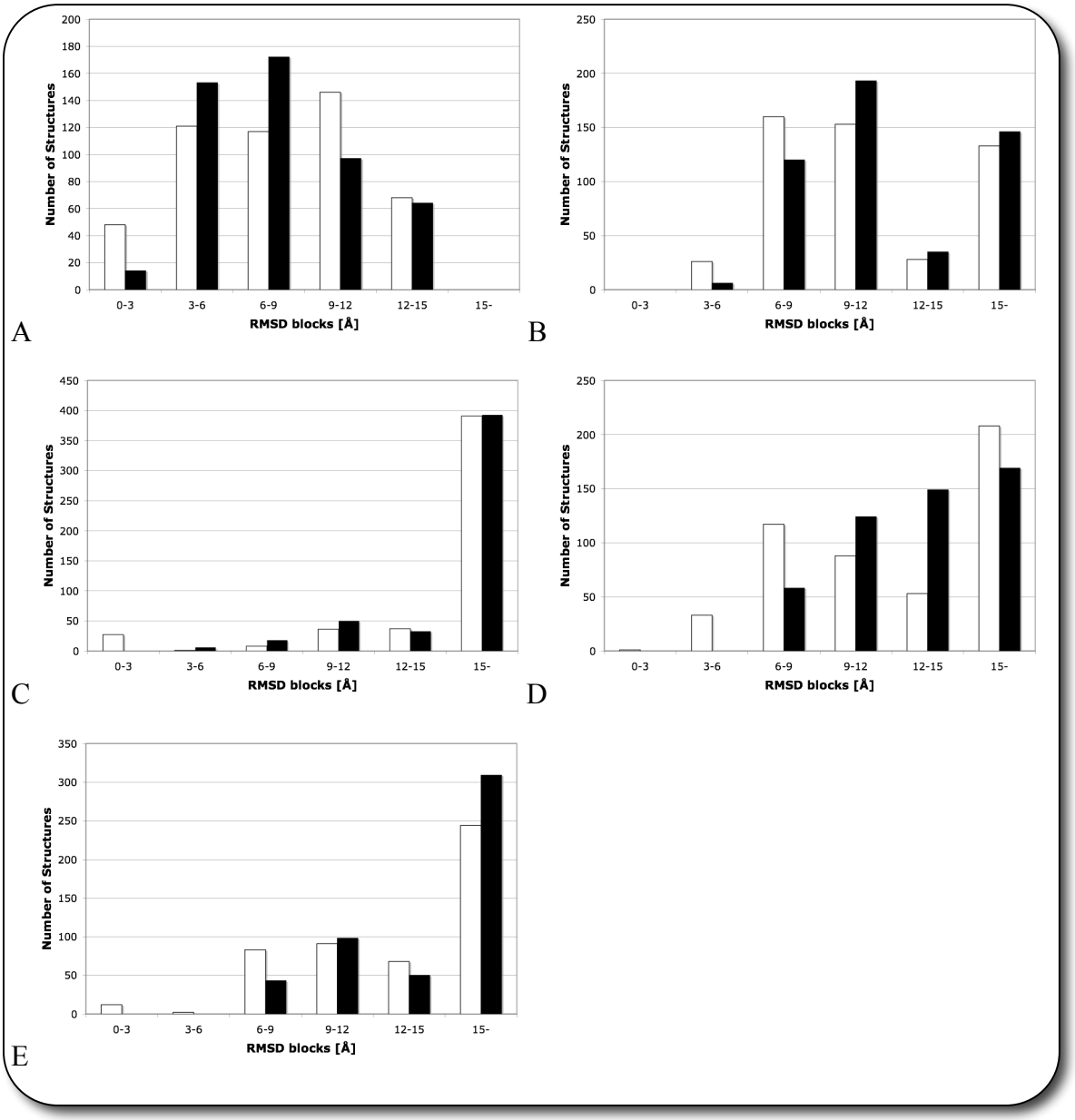


Figure 28: *Comparison of the BDOCK results based on unbound-unbound (black) and bound-bound (white) docking benchmarks. The top 500 docking results after their geometric complementarity at the interface were divided into 6 groups of RMSD values to the native complex structure. (A) 1PPE, (B) 1EWY, (C) 1F34, (D) 1D6R, (E) 1E6E.*

3.3.2. BDOCK Sensitivity without Scoring Functions

In the previous section, a poorer sensitivity of BDOCK was observed for benchmark samples with increasing conformational changes upon complex formation. In this section the overall predictability of BDOCK is analyzed for all 17 benchmark samples in the unbound-unbound docking. The results are shown in table 6. A perfect docking approach would have ranked the lowest RMSD of the docking sample to the native structure in the first position. Although even the best docking sample will not achieve a lower RMSD than the values in table 5, the ranking will most likely indicate its sensitivity. However, table 6 shows underwhelming results. The best RMSDs are mostly ranked badly once shape complementarity is used. 2PCC leads to the best results, whereas all remaining benchmarks more or less strongly vary within the top 2000 ranked structures. Some benchmarks did not even generate structures below 4Å RMSD from the native complex (average ranks for RMSD values below 4Å of table 6 with "NA"). The distribution of the results also shows no relation between difficulty of the benchmark samples judged on their conformational changes upon complex formation and the best RMSD ranking result. A more detailed aspect for the best and worst benchmark results: 2PCC and 1D6R is shown in figure 29. The graphs show that neither 2PCC nor 1D6R have any correlation between shape complementarity and near native structure prediction. This is illustrated for all benchmark samples in figure 30, where the correlation coefficients for the ranks after the docking score and the RMSD ranks to the structure of the native complex were calculated and plotted. Clearly, all computed correlations show a random relation between the docking rank and the RMSD value. In the best case, a decreasing docking rank would as well lead to a decreasing RMSD rank for the docking samples. This concludes that the sensitivity of BDOCK purely based on geometric complementarity does not lead to satisfying results in this test and strongly urges the application of a second layer of scoring functions considering not only geometry but other aspects as well. Yet, a logarithmic trend graph that was calculated for the computed correlation coefficients weakly reveals that with decreasing conformational changes the correlation coefficient between docking rank and RMSD rank increases.

Benchmark sample	Rank of the lowest RMSD	Average rank for RMSDs below 4Å	Standard deviation of average rank for RMSDs below 4Å
2SNI*	1453	1009	466.17
2SIC*	1697	902.75	594.95
2PCC*	46	45.5	0.71
2MTA	1272	**NA**	**NA**
1PPE	1371	925.92	591.16
1AY7*	1176	**NA**	**NA**
1EAW*	455	616.87	489.64
1MAH*	1877	**NA**	**NA**
1EWY	1806	1379.11	570.95
1UDI*	1849	704.77	529.80
1F34	569	**NA**	**NA**
1DFJ*	531	**NA**	**NA**
1D6R	1912	**NA**	**NA**
1E6E	951	1119.2	170.77
1HIA*	1548	858.29	778.76
1CGI*	1016	961.75	417.79
1ACB*	1370	1348	31.11

Table 6: *Benchmark results for BDOCK based on pure shape complementarity ranking. NA stands for Not Available values. *Data included in the dataset of 170 transient complexes.*

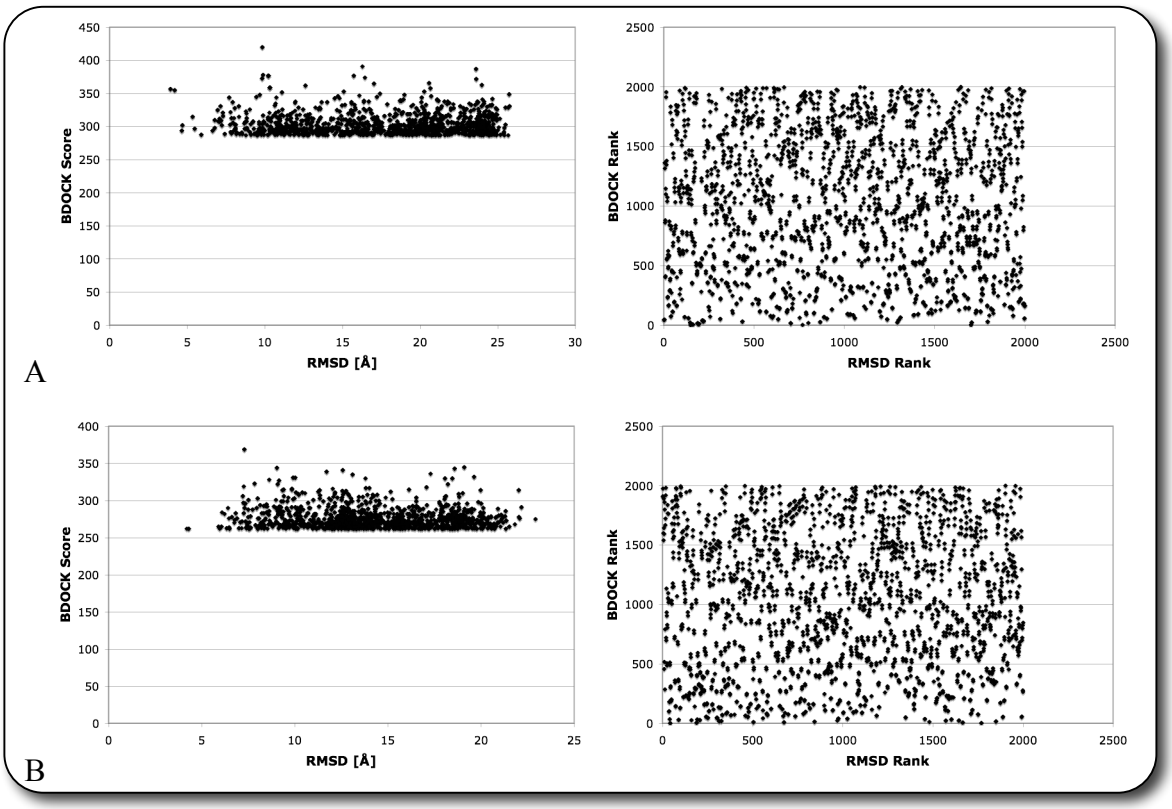


Figure 29: *Correlation analysis of the benchmark samples 2PCC (A) and 1D6R (B).*

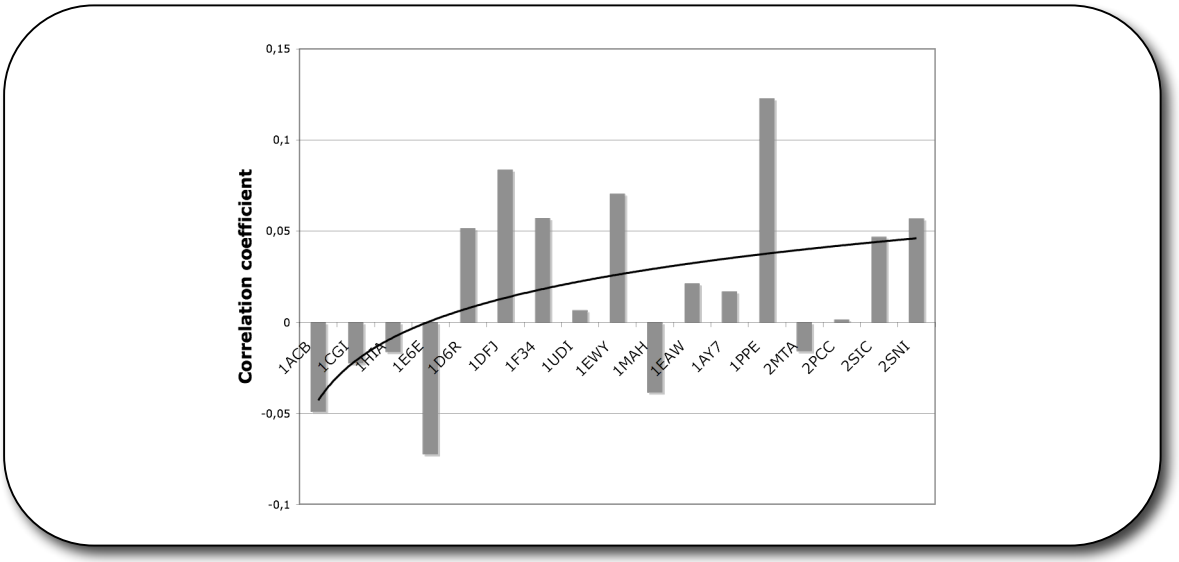


Figure 30: *Correlation coefficient of the BDOCK ranking and the native structure RMSD for given benchmark PDBs. Benchmarks are sorted from left to right by their decreasing conformational changes upon complex formation. The best 2000 structures, after their docking score, were considered in this analysis.*

3.3.3. BDOCK and SARScore(res)

In this section, the predictability of the residue-pair potential unit of SARScore (SARScore(res)) is compared to the previous results of the pure geometric scoring by BDOCK. The data for the residue-pair potentials are based on 170 transient interfaces using a distance cutoff of 5Å. Table 7 compares the results from table 6 with those of SARScore(res). In most of the cases, the best RMSD rank is lower in case for SARScore(res) when compared to the pure shape complementarity rankings of BDOCK. Focusing on those benchmark samples that were not included in the dataset of 170 transient interfaces, SARScore(res) shows more predictive results in 4 of 6 cases. This suggests a slight improvement of the docking results after re-ranking the docking samples with a residue-pair potential based on transient interfaces. Figure 31 shows the correlation coefficients of the sample ranks and RMSD ranks. The results of figure 30 are also shown. Clearly, the overall correlation of the SARScore(res) ranks are higher where the trend-graph even proposes a higher correlation for those difficult benchmarks with large conformational changes upon complex formation. A surprising aspect is found in the results that are almost equal for benchmarks from the dataset and those that were not

included. This might point to a large diversity of the residue-pairing propensities at the given distance criteria within the dataset of 170 transient complexes. Section 3.3.8 and 3.3.9 will address this aspect. However, correlation coefficients below 0.8 may still be based on random relations. This means that the observed results for SARScore(res) still do not show a significant sensitivity for the docking problem of rigid-body approaches.

Benchmark sample	BDOCK RotlR	SARScore RotlR	BDOCK ArfRb4	SARScore ArfTb4	BDOCK SdarRb4	SARScore SdarRb4
2SNI*	1453	**1324**	1009	1162.25	466.17	463.35
2SIC*	1697	**1079**	902.75	1432.75	594.95	555.08
2PCC*	**46**	552	45,5	937.5	0.71	545.18
2MTA	**1272**	1412	**NA**	**NA**	**NA**	**NA**
1PPE	1371	**178**	925.92	846.88	591.16	502.55
1AY7*	1176	**855**	**NA**	**NA**	**NA**	**NA**
1EAW*	**455**	574	616.87	945.5	489.64	460.95
1MAH*	1877	**718**	**NA**	**NA**	**NA**	**NA**
1EWY	1806	**1328**	1379.11	1474.21	570.95	278.28
1UDI*	1849	**1617**	704.77	1009.85	529.8	480.78
1F34	**569**	1326	**NA**	**NA**	**NA**	**NA**
1DFJ*	**531**	667	**NA**	**NA**	**NA**	**NA**
1D6R	1912	**763**	**NA**	**NA**	**NA**	**NA**
1E6E	951	**718**	1119.2	685.4	170.77	58.01
1HIA*	1548	**379**	858.29	406.93	778.76	94.29
1CGI*	1016	**489**	961.75	526	417.79	282.19
1ACB*	1370	**390**	1348	866.5	31.11	673.87

Table 7: *Benchmark results for BDOCK based on pure shape complementarity ranking and SARScore(res). NA stands for Not Available values, RotlR=Rank of the lower RMSD sample, ArfRb4=Average rank for RMSD below 4Å, and SdarRb4=Standard deviation of the average rank for RMSD below 4Å. *Data included in the dataset of 170 transient complexes.*

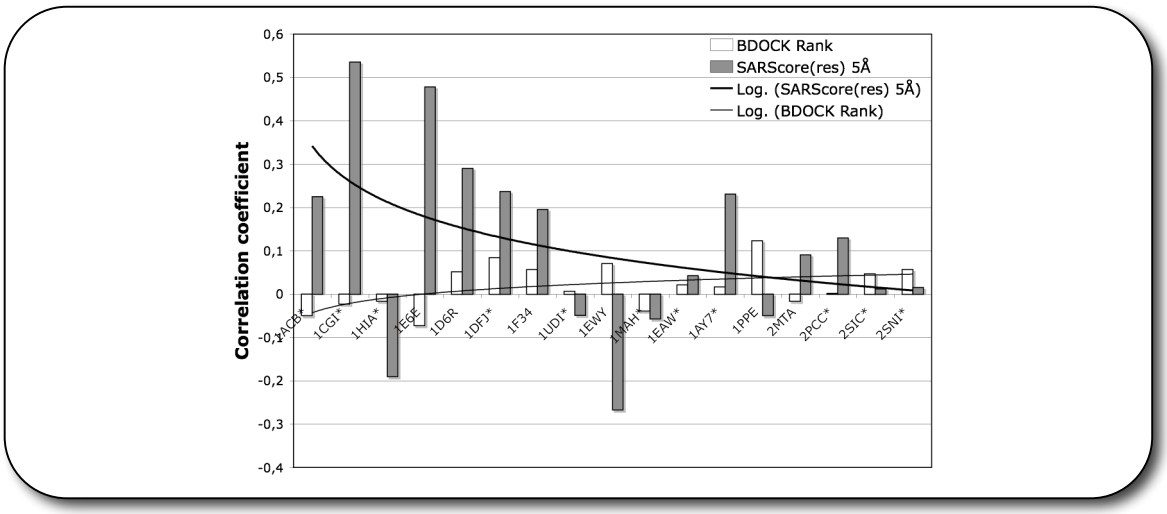


Figure 31: *Correlation coefficient of the BDOCK ranking (white)/SARScore residue pair potential at 5Å (grey) and the native structure RMSD for given benchmark PDBs. Benchmarks are sorted from left to right by their decreasing conformational changes upon complex formation. The best 2000 structures after their docking score were considered in this analysis. *Data included in the dataset of 170 transient complexes.*

3.3.4. Comparing RPScore and SARScore(res)

The previous results suggest a slightly higher predictability of the residue-pair potential SARScore(res). The question addressed in this section is whether the residue-pair potential based on a more suitable dataset (here SARScore(res) compared to RPScore) will lead to more sensitive results. Although the available RPScore potentials specify a distance cutoff of 5Å it is not clear whether this cutoff is used on the Cβ atoms, any atom, or side-chain atoms. Since the authors could not supply this missing information, the following results should be treated carefully. Table 8 and figure 32 illustrate the results. Apparently, the sensitivities of RPScore and SARScore(res) are nearly the same in table 8. This is the case for all benchmarks as well as for those that were not included in the dataset of 170 transient interfaces. As the authors did not mention the detailed interface criteria, no information on the 103 non-homologous data of RPScore could be retrieved as well, so the results could be focused on those benchmark samples that have not been used in any of the residue-pair potential training sets. When analyzing the correlation coefficients for residue-pair rank and RMSD rank, SARScore(res) appears more sensitive showing higher correlations, aside the ambiguity on the comparability of the two

potentials as they may be based on different interface criteria. This is mainly the case for the rather difficult cases.

Summarizing the overall results by also taking into account the trend graphs gives the SARScore residue potential a slight advantage over RPScore when considering the predictability of the used benchmark samples. Yet, the achieved correlation coefficients are too low to clearly prove the assumption that more tailored scoring matrices will lead to better docking scoring-functions rather than all-purpose solutions such as RPScore.

Benchmark sample	RPScore RotlR	SARScore RotlR	RPScore ArfRb4	SARScore ArfTb4	RPScore SdarRb4	SARScore SdarRb4
2SNI*	**301**	1324	917.25	1162.25	566.61	463.35
2SIC*	**89**	1079	517	1432.75	448.92	555.08
2PCC*	1649	**552**	1793.5	937.5	204.35	545.18
2MTA	**535**	1412	**NA**	**NA**	**NA**	NA
1PPE	663	**178**	842.97	846.88	454.56	502.55
1AY7*	**491**	855	**NA**	**NA**	**NA**	**NA**
1EAW*	869	**574**	996	945.5	374.99	460.95
1MAH*	**213**	718	**NA**	**NA**	**NA**	**NA**
1EWY	**149**	1328	527.11	1474.21	515.02	278.28
1UDI*	**317**	1617	635.15	1009.85	386.99	480.78
1F34	**465**	1326	**NA**	**NA**	**NA**	**NA**
1DFJ*	**259**	667	**NA**	**NA**	**NA**	**NA**
1D6R	1373	**763**	**NA**	**NA**	**NA**	**NA**
1E6E	1361	**718**	1346.8	685.4	459.14	58.01
1HIA*	1163	**379**	1483.36	406.93	278.09	94.29
1CGI*	930	**489**	1294.87	526	355.99	282.19
1ACB*	1009	**390**	1052.5	866.5	61.52	673.87

Table 8: *Benchmark results for two residue based pair potentials: RPScore and SARScore(res). NA stands for Not Available values, RotlR = Rank of the lower RMSD sample, ArfRb4 = Average rank for RMSD below 4Å, and SdarRb4 = Standard deviation of the average rank for RMSD below 4Å. *Data included in the dataset of 170 transient complexes.*

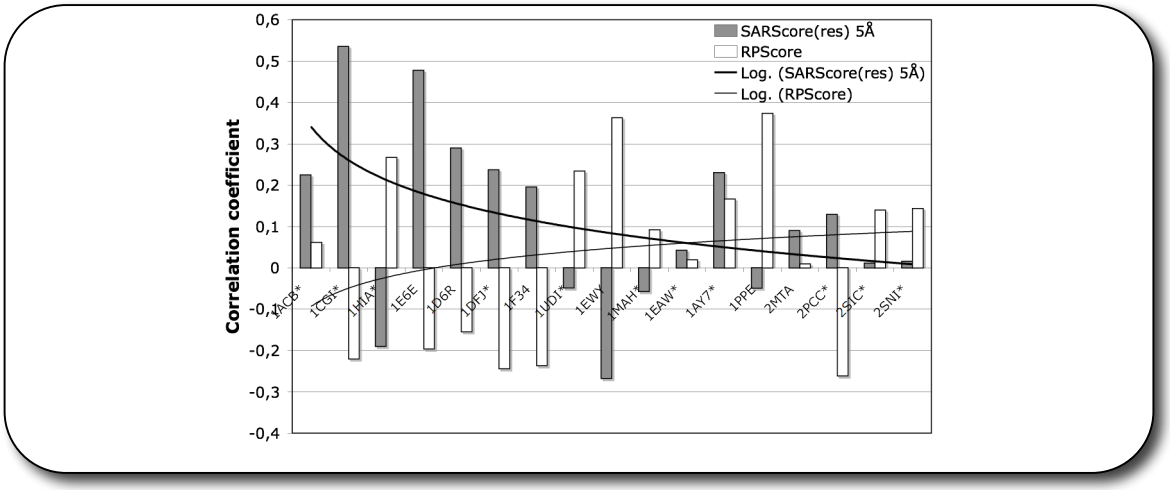


Figure 32: *Comparing correlations between RPScore and SARScore residue pair potential rankings and the RMSD ranks. *Data included in the dataset of 170 transient complexes.*

3.3.5. BDOCK and SARScore(struc)

In section 3.3.3 the pure shape complementarity of BDOCK was compared to the residue-pair potential of SARScore. Here, the secondary structure element-pair potentials of SARScore (SARScore(struc)) are compared to BDOCKs geometric evaluation of docking samples. Table 9 reveals a similar performance for SARScore(struc) as it was the case for SARScore(res) in table 7. Figure 33 leads to a lower trend graph once it is compared to the residue-pair potential of SARScore. This is also underlined in table 10, where the results of SARScore(res) are directly compared to those of SARScore(struc).

Benchmark sample	BDOCK RotlR	SARScore RotlR	BDOCK ArfRb4	SARScore ArfTb4	BDOCK SdarRb4	SARScore SdarRb4
2SNI*	1453	**836**	1009	1127.62	466.17	639.29
2SIC*	1697	**1150**	902.75	1274.25	594.95	392.09
2PCC*	**46**	814	45,5	951.5	0.71	194.45
2MTA	1272	**1161**	**NA**	**NA**	**NA**	**NA**
1PPE	1371	**607**	925.92	942.6	591.16	425.75
1AY7*	1176	**920**	**NA**	**NA**	**NA**	**NA**
1EAW*	**455**	605	616.87	843	489.64	406.29
1MAH*	1877	**1222**	**NA**	**NA**	**NA**	**NA**
1EWY	1806	**1357**	1379.11	1423.58	570.95	238.86
1UDI*	1849	**946**	704.77	699.23	529.8	452.7
1F34	**569**	818	**NA**	**NA**	**NA**	**NA**
1DFJ*	**531**	1353	**NA**	**NA**	**NA**	**NA**
1D6R	1912	**828**	**NA**	**NA**	**NA**	**NA**
1E6E	**951**	1529	1119.2	1244.6	170.77	418.52
1HIA*	1548	**682**	858.29	544.79	778.76	203.52
1CGI*	1016	**421**	961.75	384.5	417.79	229.35
1ACB*	1370	**201**	1348	533.5	31.11	470.23

Table 9: *Benchmark results for BDOCK based on pure shape complementarity ranking and SARScore(struc). NA stands for Not Available values, RotlR = Rank of the lower RMSD sample, ArfRb4 = Average rank for RMSD below 4Å, and SdarRb4 = Standard deviation of the average rank for RMSD below 4Å. *Data included in the dataset of 170 transient complexes.*

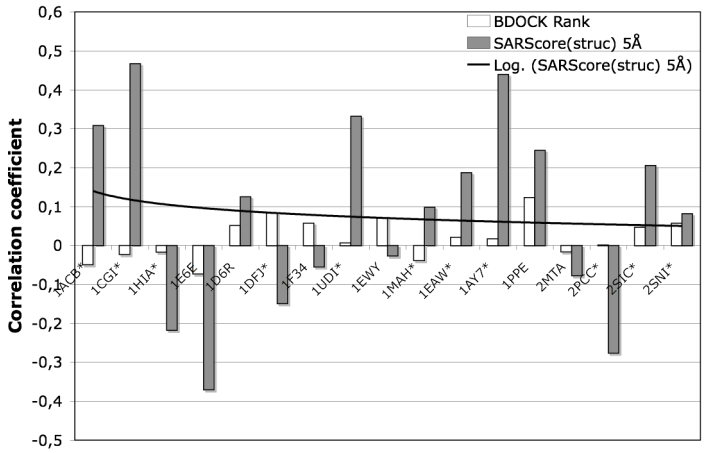


Figure 33: *Correlation coefficient of the BDOCK ranking (white)/SARScore secondary structure element-pair potential at 5Å (grey) and the native structure RMSD for given benchmark PDBs. Benchmarks are sorted from left to right by their decreasing conformational changes upon complex formation. The best 2000 structures after their docking score were considered in this analysis. *Data included in the dataset of 170 transient complexes.*

Benchmark sample	Res RotlR	Struc RotlR	Res ArfRb4	Struc ArfTb4	Res SdarRb4	Struc SdarRb4
2SNI*	1324	**836**	1162.25	1127.62	463.35	639.29
2SIC*	**1079**	1150	1432.75	1274.25	555.08	392.09
2PCC*	**552**	814	937.5	951.5	545.18	194.45
2MTA	1412	**1161**	**NA**	**NA**	NA	**NA**
1PPE	**178**	607	846.88	942.6	502.55	425.75
1AY7*	**855**	920	**NA**	**NA**	**NA**	**NA**
1EAW*	**574**	605	945.5	843	460.95	406.29
1MAH*	**718**	1222	**NA**	**NA**	**NA**	**NA**
1EWY	**1328**	1357	1474.21	1423.58	278.28	238.86
1UDI*	1617	**946**	1009.85	699.23	480.78	452.7
1F34	1326	**818**	**NA**	**NA**	**NA**	**NA**
1DFJ*	**667**	1353	**NA**	**NA**	**NA**	**NA**
1D6R	**763**	828	**NA**	**NA**	**NA**	**NA**
1E6E	**718**	1529	685.4	1244.6	58.01	418.52
1HIA*	**379**	682	406.93	544.79	94.29	203.52
1CGI*	489	**421**	526	384.5	282.19	229.35
1ACB*	390	**201**	866.5	533.5	673.87	470.23

Table 10*: Benchmark results for residue (res) and secondary structure element (struc) - pair potentials of SARScore. NA stands for Not Available values, RotlR = Rank of the lower RMSD sample, ArfRb4 = Average rank for RMSD below 4Å, and SdarRb4 = Standard deviation of the average rank for RMSD below 4Å. *Data included in the dataset of 170 transient complexes.*

3.3.6. Critical Assessment of the Results

The previous results remain unsatisfactory. None of the methods achieved acceptable prediction accuracies for the given benchmark set. Although the trend shows a weak advantage for the SARScore residue-pair potential, all correlation coefficients are yet in the area of random distributions. At this point, estimating the performance of a random residue and secondary structure element-pair potential seems helpful. Therefore a residue and secondary structure element-pair potential matrix was randomly generated and used for re-ranking the top 2000 docking samples derived from BDOCK. Figure 34 shows for each scoring unit the average correlation coefficient over all 17 benchmark samples. Previously it was found that benchmarks from the dataset do not necessarily lead to better results. Thus, averaging the scores over all benchmark samples seems eligible. Although

the randomly generated secondary structure element-pair potentials result in the lowest average correlation, the randomly generated residue-pair potentials score surprisingly high. This supports the apprehension that the overall results may hardly be better than random.

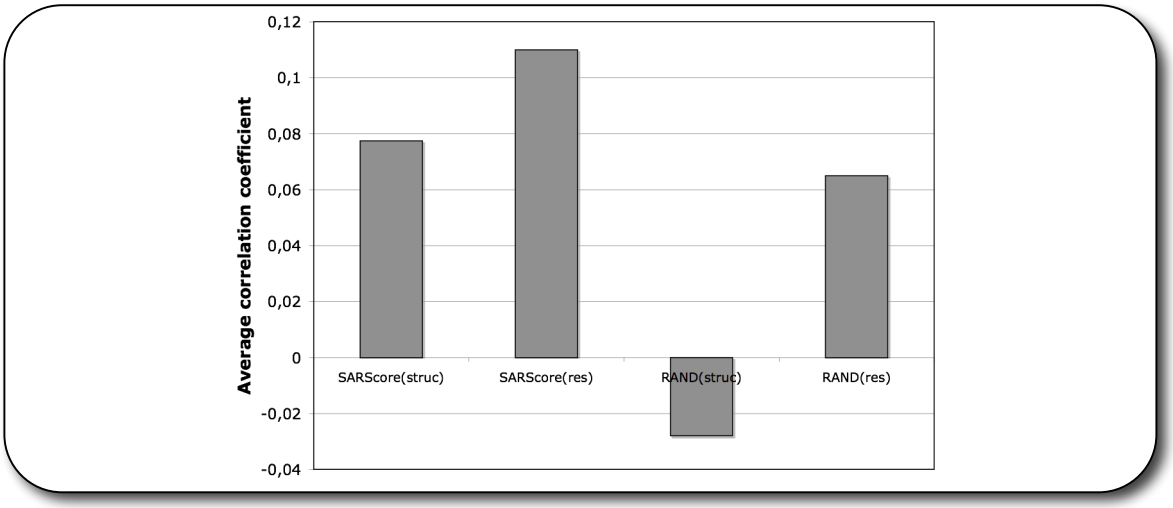


Figure 34: *Average correlation coefficients SARScore(struc) and SARScore(res) are compared against random potentials RAND(struc) and RAND(res).*

3.3.7. Analysis of the Distance Criteria

As the definition of the interface area definitively describes the pair potentials and the previous results for the distance criterion of 5Å did not lead to clear observations, analyzing different distance criteria was another choice. Here, the distance cutoffs of 4Å, 5Å, 6Å, and 7Å were compared to each other. Figure 35 compares the average correlation coefficients over all 17 benchmark samples. Apparently, the larger the distance cutoff is set, the higher the average correlation and the predictability become. This is surprising since a distance cutoff of 7Å between any heavy atoms on two interacting chains may statistically evaluate a large number of non-interacting residues. However, the average scores are still much too low. Figure 36 shows the correlation coefficients for each benchmark sample separately. A drastic improvement of the larger distance cutoffs compared to smaller ones is not found and the small improvement of larger distance cutoffs is on average negligible.

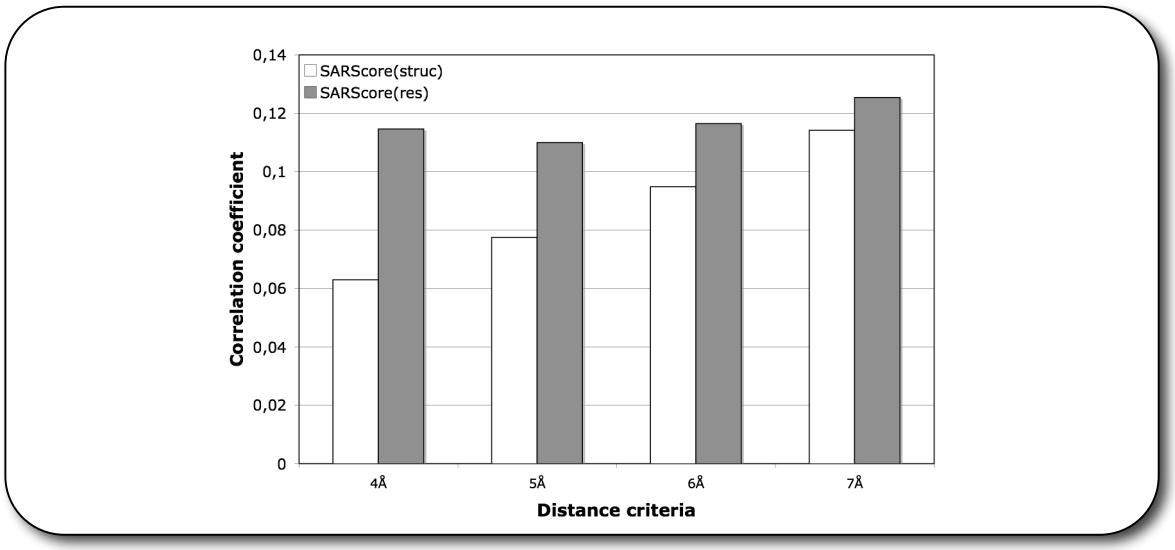


Figure 35: *Comparing the average correlation coefficients of the benchmark set for given pair potentials depending on the distance cutoff for the interface area.*

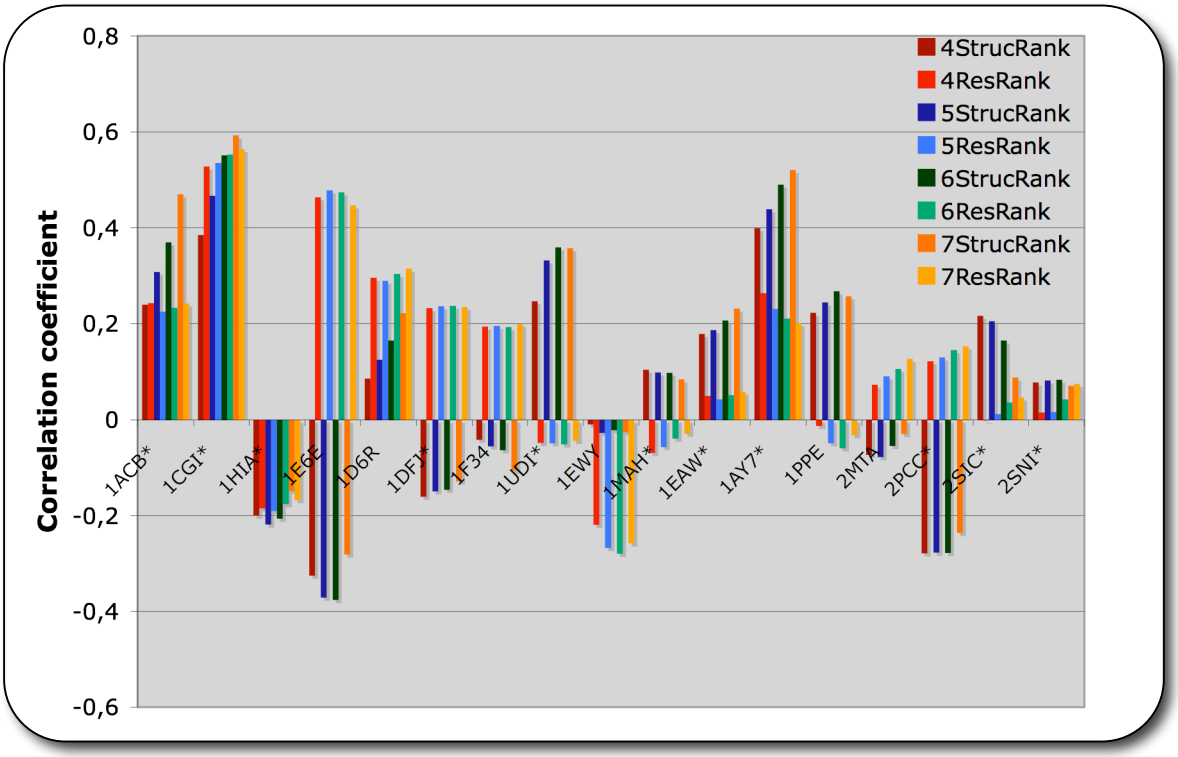


Figure 36: *Comparing the correlation coefficients for given pair potentials depending on the distance cutoff for the interface region. 4StrucRank stands for SARScore(struc) derived from interface data based on a distance cutoff of 4Å. "Rank" indicates the correlation coefficients based on rank distributions instead of score values. *Data included in the dataset of 170 transient complexes.*

3.3.8. Analysis of the Benchmark Set

Summing up most of the previous results leads to mixed information. Mainly the inhomogeneous results in the scoring do not allow any conclusions. It was noticed that the rather large angle steps for the rotations of the mobile protomer may lead to rather bad docking samples, which may become an unsuitable basis for the scoring function as well. Smaller angle steps most likely may lead to better docking samples but the increase of computational time would be tremendous. Given this limitation, other questions arise: Are residue-pair potentials able to predict the native complex structures at all? Do native complexes or even transient complexes have stable patterns in their residue propensities? Ofran and Rost reported clear differences within several types of interactions [17]. However, the authors used a given criteria for distinguishing between the interface types and computed average residue pair propensities. It was not evaluated whether an observed pattern is strongly conserved in all participating complexes. In this section, the question will be addressed on whether the benchmark samples that yield rather well, average, or bad results do share residue-pair propensity patterns. Figure 37 reveals a surprising finding comparing the residue propensities of all benchmark samples in the categories: good (green), average (yellow), and bad (red) retrieved from a distance cutoff of 5Å. Within each category, the correlation coefficients of the residue-pair potentials were computed in order to search for patterns that may cause the shared level of prediction accuracy based on SARScore(res). It was expected that at least those benchmark samples that were acceptably predictive might contain a similar residue-pairing propensity pattern and therefore lead to higher correlation coefficients. In figure 37 all correlations are clearly too low. Patterns cannot be found. All derived results for the order of ranks after pair-potential scoring may be purely by chance since the benchmark samples used here do not bury any patterns. For the given interface criterion and benchmark set, this therefore generally questions the usability of any residue-pair potential to enhance the sensitivity of rigid-body docking.

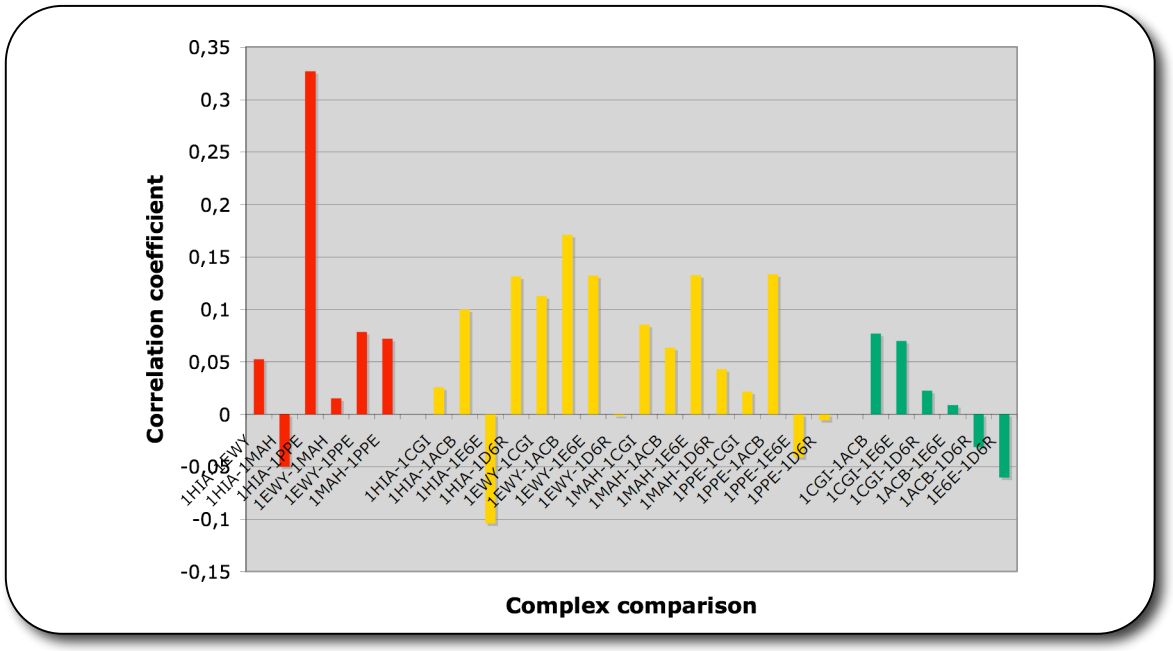


Figure 37: *Comparing residue-pairing propensities for a distance cutoff of 5Å within benchmark samples that led to sensitive scoring results for residue-pair potentials (green), average sensitive results (yellow), and low sensitive results (red).*

3.3.9. Analysis of the Dataset

The essential idea behind residue-pair potentials is the existence of patterns within a group of complexes. Ofran and Rost found clearly different residue-pairing propensity patterns within 6 types of interfaces [17]. In the dataset presented in chapter 2, 170 transient interfaces were collected. Although it was shown that the average pairing propensities are in agreement with the expected properties of transient complexes, it has not yet been evaluated how well conserved these patterns are within the 170 transient complexes. In this chapter, correlation coefficients within all complexes of the dataset are computed and clustered with the Neighbor Joining algorithm. The cluster tree is graphically shown in figure 38 drawn using the program MEGA3.1 [126]. There are a large number of clusters based on their residue-pairing propensities at a distance cutoff of 5Å. Out of 14,365 correlation coefficient calculations the weakest correlation lies at -0.16 (#17 and #102) and the highest at 0.99 (#135 and #152). The average correlation coefficient lies at 0.15 and marks a very low value and no stable pattern for all complexes. This observation may be limited to the applied distance criterion of 5Å. In the following chapters other criteria will be evaluated.

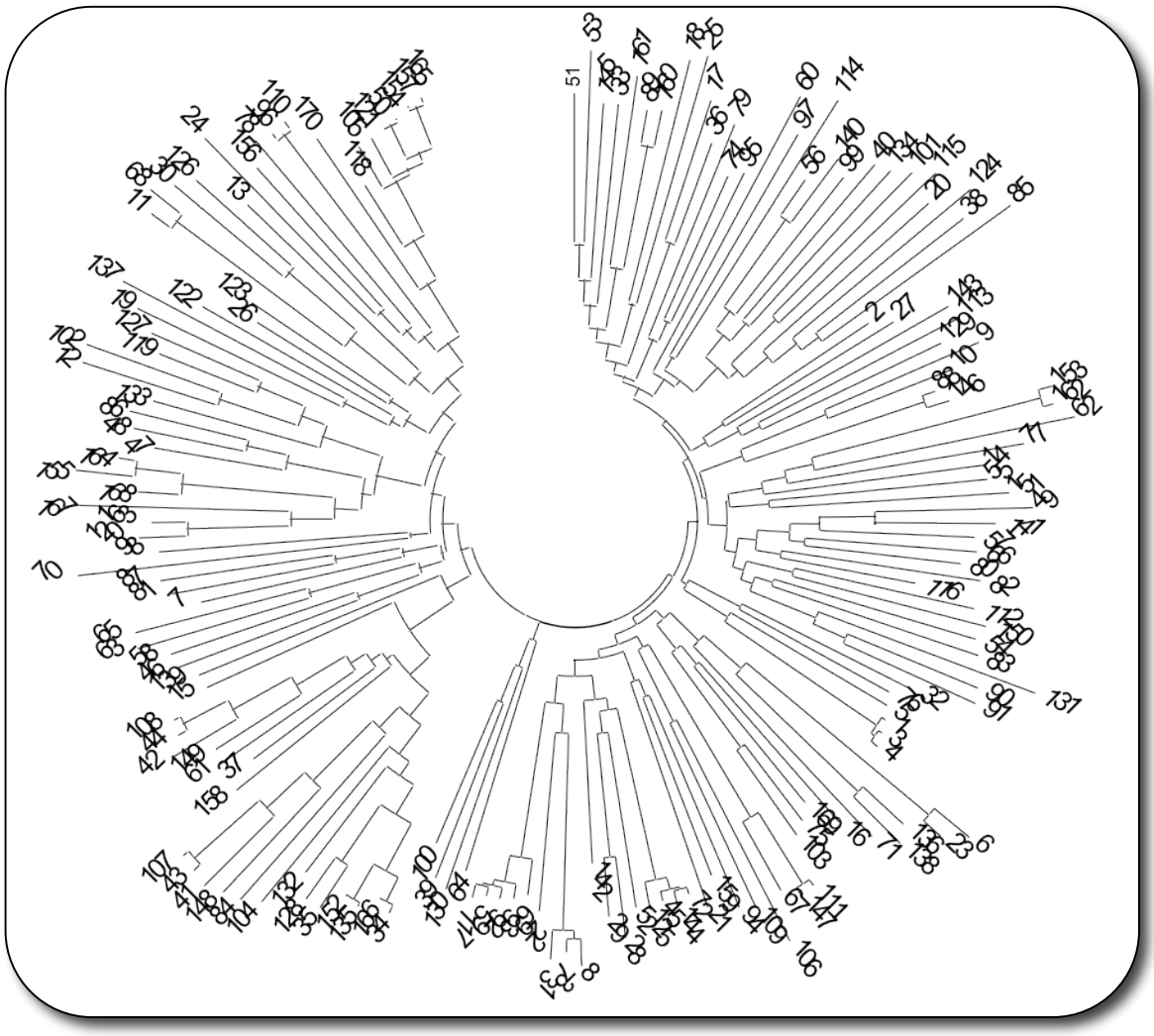


Figure 38: *Circular tree of the clustered complexes using MEGA3.1 and the Neighbor Joining algorithm. The longer the branches, the lower the correlation coefficient and thus the less similar the complexes' residue-pairing propensities at a distance cutoff of 5Å. Complexes were numbered from 1 to 170 and used as taxon for the branches. Apparently, no clear pattern exists within these transient complexes with respect to their residue-pairing propensities.*

3.4. Conclusion and Outlook

The application of rigid-body protein-protein docking is a common practice in predicting complex structure of known protomers but unknown complexed states. Such approaches are fast and allow a full conformational search. As proteins change their conformations upon complex binding, this was identified as a major weakness for rigid-body docking approaches. In this project, an implementation of FFT docking was performed in order to test this weakness and apply common attempts to enhance its sensitivity. Although BDOCK supports

partial protein flexibility by allowing surface penetrations, it was shown that docking of unbound protein conformations leads to less accurate predictions when compared to the results of bound proteins. To possibly overcome this issue and the obvious weakness of evaluating the docking conformations purely by geometric complementarity of the interface region, a subsequent scoring unit was implemented. This scoring unit loads the top 2000 docking samples ranked by their geometric interface complementarity and recalculates their ranks based on residue and secondary structure element-pair potentials. 17 benchmark samples from the ZLAB 2.0 were used to compare the docking accuracies. It was found, that neither the docking alone nor the extended scoring unit lead to satisfying accuracies in the case of the 17 benchmark samples. However, it was also shown that residue and secondary structure element-pair potentials do enhance the purely geometry based docking. Yet, these enhancements may rely on random effects. This also affects the observation that there is little improvement of the results once more suitable residue-pair potentials are employed. The dataset did not contain any clearly conserved patterns of residue-pairing propensities at a distance cutoff of 5Å as it was also the case for the benchmark set. For the given benchmark conditions, residue and secondary structure element-pair potentials will not enhance FFT docking approaches in any significant way.

A generally controversial aspect of benchmarking docking approaches lies in computing the RMSD from the full protein complex. The entire protein complex may involve very similar interface conformations but strongly different orientations over the full complex. This will result in a high RMSD even though the interaction site was predicted correctly. It was suggested to rather restrict the RMSD calculations on the interface area. However, this strategy as well buries the risk for wrong evaluations. For small interfaces on large proteins, which is more likely the case in transient complexes, two interfaces may have a low deviation but still be located on different surface patches.

As for the scoring function, different interface criteria should be tested since a correct definition of this area plays a major role in the predictability of computed potentials. Furthermore, conserved patterns within transient complexes should be analyzed based on different interface properties. For that it may be helpful to compare the set of transient interfaces with an out-group of permanent interactions to enhance the contrast in their patterns.

4 Distinction of Obligate and Non-obligate Interactions

4.1. Overview

A dataset of obligate/permanent and non-obligate/transient complexes is used to test a number of interface properties retrieved from two different interface criteria on whether the data can be clearly divided into obligate/permanent and non-obligate/transient complexes. This chapter addresses the previously posed question of available patterns at the interface regions of transient complexes for given interface criteria and interface properties. Clearer results are assumed by including an out-group of obligate/permanent interfaces into the analysis. To test the importance of certain properties of amino acids, residue classes are considered that group amino acids with similar qualities, and potentials are computed based on these assemblies. Ultimately, the complexes are clustered based on the similarity of their interfaces for a given distance criterion and interface property. To visualize the clusters, a method is applied that is mostly used in the field of phylogenetics; distances between the complexes for a given property are computed and a tree is drawn based on these distances. A statistical test leads to the best separation.
This study was done with the support of a MySQL database developed by Peter Walter in his supervised diploma-thesis completed in 2006.

4.1.1. Introduction

Zhu et al. recently published an automatic classification for distinguishing obligate and non-obligate complexes [68]. Based on a dataset of 75 obligate and 62 non-obligate complexes, the authors achieved a separation accuracy of 91.8% when combining three of the six interface properties, namely the absolute and relative interface area sizes and the amino acid-composition of the interface area-normalized. The same database is used in this chapter in order to test the separation quality of 19 different interaction-site

properties. Computed correlation coefficients between each combination of the complexes were then converted into distances. Applying the program package MEGA3.1 [126] including 3 different clustering algorithms, two groups of clusters were calculated as the dataset consists of obligate and non-obligate complexes. The χ^2 test was applied on the clustered groups estimating the statistical significance for a given interface property to properly cluster the dataset into obligate and non-obligate complexes. The 19 interaction-site properties consist of 8 different residue compositions and 8 residue-pairing propensities, two different secondary structure element-pairing propensities, and the tightness of the interaction site. Due to the large number of features (20 for residue-composition data and 210 for residue-pairing propensities) a graphical output of the clusters is nearly impossible. However, using MEGA3.1 and correlation coefficients, a dendrogram was found very useful for visualizing the clusters.

4.2. Methods

4.2.1. Data Handling

The cited work of Zhu et al. presents a new web-application called NOXclass [68]. It is an automated classifier for distinguishing obligate, non-obligate and crystal packing interactions. Beside 106 crystal packing contacts the authors also collected 75 obligate and 62 non-obligate complexes in order to train their system. Obligate and non-obligate interactions were taken from a compiled set from Bradford et al. [127]. A set of transient interactions [128] was added to the non-obligate data, which share the same definition. Exactly this dataset of Zhu et al. was used in this chapter. Although 170 transient/non-obligate interactions were collected in chapter 2, they were not included in this analysis as they may shift the rather balanced rate of obligate and non-obligate interactions from Zhu et al.

In this study, the definition for interface participating residues was chosen differently from Zhu et al., based on experiences from available crystal structures where interfaces may involve tightly bound areas but also complementary surface patches separated by one or more water layers [25]. Zhu et al. defined a residue as an interface residue once its

solvent accessible-surface area (SASA) decreased by more than 1Å^2 upon formation of the complex. In previous studies, a distance criterion of 5Å between heavy atoms was used to retrieve interface atoms, interface residues, and interface secondary structure elements. Here, two extreme distance-cutoffs are applied: 4Å and 8Å. Comparing results for distance cutoffs of 4Å and 8Å helps identifying the most suitable criteria for an interface and simplify the distinction of obligate and non-obligate interfaces. While 4Å considers only very tightly bound regions of the interface, barely allowing water penetration, a distance-range of 8Å even includes peripheral electrostatic interactions. However, once a distance cutoff of 8Å is used, one faces the difficulty of considering many buried residues as well that most likely do not participate in the interaction. To avoid such buried residues to be counted as interface residues, a new criterion for interface residues was added, where a residue fulfilling the distance criterion also has to have a larger surface contribution than 0Å^2 when a probe with a radius of 4Å is used to calculate the surface area. Although common SASA calculations use a probe size of 1.4Å, which is a typical radius of a water molecule, a preliminary test showed some peculiarities resulting from this probe size. Based on the current dataset, the test with a VMD-script showed that at the probe size of 1.4Å many buried cavities are counted in the surface area and thus many buried residues are wrongly predicted as surface exposed. Table 11 shows the ratio of buried interface residues for varying probe radii based on the SASA measurements of the program package VMD [65] (also see section 1.2.4.1). Clearly, the ratio for probe sizes up to 3.0Å is much lower than expected. Only for r = 4.0Å, a ratio of 18% was obtained, which fits the expected range. Therefore a probe size of 4Å was used, which suppresses any cavities smaller than 268Å^3 size. This approximately equals the volume of 9 water molecules.

Probe radius [Å]	Rate of buried interface residues at 8Å distance-cutoff [%]
1.4	0
2.4	0.05
3.0	1.59
4.0	18.54

Table 11: *Ratio of buried potential interface residues at a distance-cutoff of 8Å for different surface probe radii. 22869 residues in 137 complexes were examined.*

Similar to previous works, statistical information on the residue composition, residue-pairing propensities and secondary structure element-pairing propensities were collected using a tcl/tk script and the program package VMD. In addition to the secondary structure element-pairing propensities, another yet simple criterion for the tightness of the interaction site was introduced. Using the two counts of residue pairs within the distance cutoff of 4Å and 8Å, the tightness of the fit was simply defined as the ratio of the two numbers. A tight complex will have a ratio close to one. Lower values indicate less tight contacts, where the number of interactions at 8Å distance cutoff greatly exceeds the number of 4Å interactions.

Collecting these data yielded a large set of information. To facilitate analysis, Walter and coworkers created a relational database using open-source components only (see chapter 5). This system allows quick handling of the large amount of data and an easy import of additional data. Another feature of the database is the grouping of residues into classes of amino acids with similar properties (e.g. hydrophobic and hydrophilic). As shown in table 12, the simplest classification separates all amino acid-types into H-bond forming and non-forming residues (group label 2). The next finer level accounts for the different physicochemical properties and divides the amino acids into hydrophobic, hydrophilic uncharged, negatively and positively charged residues (group label 4). The last level contains small, hydrophobic, negatively and positively charged, and polar amino acids (group label 5). Grouping amino acids into classes should allow distinguishing important mutations from mutations maintaining the same properties.

Group 5	Property	small	hydrophobic	negatively charged	positively charged	polar
	Amino acids	AGPST	CLIVMWFY	ED	KRH	NQ
Group 4	Property	hydrophobic	Hydrophilic uncharged	negatively charged	positively charged	
	Amino acids	AVLIFMGWP	STCNQHY	ED	KR	
Group 2	Property	Not H-bond forming		H-bond forming		
	Amino acids	AVLIFMGWP		STCNQHYEDKR		

Table 12: *Amino acid-assemblies for residue composition and residue-residue pairing propensities. Amino acid-names are abbreviated as one-letter code.*

Due to the unequal abundances of the residue distributions upon assembling, a simple fraction method based on the number of amino acids in each group was introduced. The score of an amino acid-assembly composition or pairing propensity-group was calculated from the logarithmic ratio of counted and expected value, as it was applied with the mole-fraction and contact-fraction method in section 3.2.2.1. The expected value was derived from the number of amino acids in that group, e.g. 5/20 in the group 5 for "small". Additionally, area normalization for the residue composition and residue-pairing propensity data in all amino acid-classes was used. Therefore, the SASA at a probe size of 1.4Å was calculated for each interface residue. Dividing the SASA contribution of each individual interface residue at a given position by the total size of the interface, a relative surface contribution for each interface residue could be calculated. Other previously used fraction methods such as the contact-fraction and the mole-fraction have a constant effect on all complexes with only different fractions for different types of amino acids. Since the differences among the complexes should be emphasized, such fraction methods will not influence the correlation coefficients between the complex properties.

For the residue compositions the database outputs a table containing 137 lines, where each line includes residue composition of each complex for all 20 amino acids (20 columns). Applying the amino acid-classes to this table resulted in three further tables of 137 lines each and – depending on the amino acid-class – 5, 4, or 2 columns. Given the two distance criteria, the residue-composition property results in 8 different tables. Similar to this, the residue-pairing propensity tables from the database contain 137 lines and 210, 15, 10, and 3 columns in two times four tables. The secondary structure element-pairing propensities were not considered in any classes and therefore yield two tables with 137 lines and four columns (helix, beta, turn, and coil). Finally, one table was computed for the tightness of the fit, which was derived from the ratio of the number of interface residues at 4Å and 8Å at a given complex. This table consists of 137 lines and one column.

4.2.2. Distance Matrix

Similarity distances were computed from Pearsons' correlation coefficients of 19 different interface properties within 137 complexes:

$$c_{ij} = \frac{\sum ij - \frac{\sum i \sum j}{N}}{\sqrt{\left(\sum i^2 - \frac{\left(\sum i\right)^2}{N}\right)\left(\sum j^2 - \frac{\left(\sum j\right)^2}{N}\right)}}$$

where c_{ij} is the correlation coefficient for a given property of complexes i and j and N the number of elements for this interface property (i.e. N = 400 for the residue-residue pairing preferences in 20x20). To avoid negative similarity values the correlation coefficients c_{ij} were converted into positive distances:

$$d_{ij} = 10 - 10c_{ij}$$

where d_{ij} is the distance score for the correlation coefficient c_{ij} with a range of 0-20. Zero indicates the lowest possible distance with the highest correlation coefficient of 1. In the case of interface tightness, the ratio was directly scaled to a range of 0-20 without calculating the correlation coefficient. 19 pair distance-lists with 9316 ($^{1}_{137}$) distance pairs each were generated. In order to import the pair-distance table into MEGA3.1 the table was converted into an upper-right matrix. Finally, the following clustering algorithms implemented in MEGA3.1 were used: neighbor joining (NJ), minimum evolution (ME) and unweighted pair group method with arithmetic mean (UPGMA) (see also section 1.2.5.1).

4.2.3. Significance Assessment

The major focus of this work is to find the interface properties that lead to significant separations of obligate and non-obligate complexes and reveal conserved patterns of the given interface property. Judging on the results after the first bifurcation (assumed separation into obligate and non-obligate interactions) of each clustering algorithm leads to a 2x2 matrix for each clustering algorithm and interface property where the

distribution of obligate and non-obligate complexes in each branch is statistically analyzed. To estimate the significance of this clustering the Pearson's χ^2-test including the Yates correction and the associated p value is employed:

$$\chi_c^2 = \frac{\left(\left| n_{1obl} \times n_{2non} - n_{2obl} \times n_{1non} \right| - \frac{n_{all}}{2} \right)^2 \times n_{all}}{n_{obl} \times n_{non} \times n_1 \times n_2}$$

where χ_c^2 is the corrected χ^2 after Yates, n_1 and n_2 the first two branches after the first bifurcation (see figure 41) and n_{obl} and n_{non} the number of obligate and non-obligate complexes. Corresponding to this, n_{1obl} is the number of obligate structures in the first branch and n_{1non} the number of non-obligate structures in the same branch. The same case is for n_{2obl} and n_{2non} in the second branch. n_{all} is the sum of n_{1obl}, n_{1non}, n_{2obl} and n_{2non} and equals 137 in this work.

4.3. Results

The focus of this analysis lies on the interface properties that are leading to the highest χ_c^2 values and thus clearest separation between obligate and non-obligate complexes. Finding suitable interface properties that show conserved patterns within obligate and non-obligate complexes will allow a more enhanced sensitivity in scoring rigid-body docking approaches. Furthermore, this knowledge may facilitate the generation of larger databases and therefore enhance the statistical strength of the data.

4.3.1. Evaluating the Clusters for given Properties

Figure 39A shows the χ_c^2 values for all 19 features. Additionally, 2 models are shown that contain a perfect and a random separation. Three criteria achieved χ_c^2 values of more than 15 at the distance cutoff of 8Å. The classification by residue-pairing propensities scored in a χ_c^2 value of 15.03 in the case of pairing propensities of H-bond forming residues and those that do not form such bonds. Taking a closer look at the average scores for retrieving the pair distance-lists of the pairing propensities reveals a trend

shown in table 13A. The unfavored interaction between H-bond forming and non-forming residues marks an unexpected high average score. This indicates the problem of a distance-cutoff of 8Å where many unspecific contacts between accessible atoms are counted. However, since these unspecific contacts do not differ significantly between the obligate and non-obligate data set, the focus stays on the interactions of H-bond forming and H-bond non-forming groups. The different properties are obvious here. Obligate complexes have much fewer H-bond forming interactions than non-obligate complexes. These results are strongly related to another high scoring χ_c^2 value of figure 39A: the H-bond forming and non-forming composition at the interfaces, which achieved a χ_c^2 value of 16.2. Table 13B shows the average scores for obligate and non-obligate complexes. This distribution is in full agreement with the previous results and once more underlines the importance of the capability to form or not form H-bonds once it comes to the classification of obligate and non-obligate complexes.

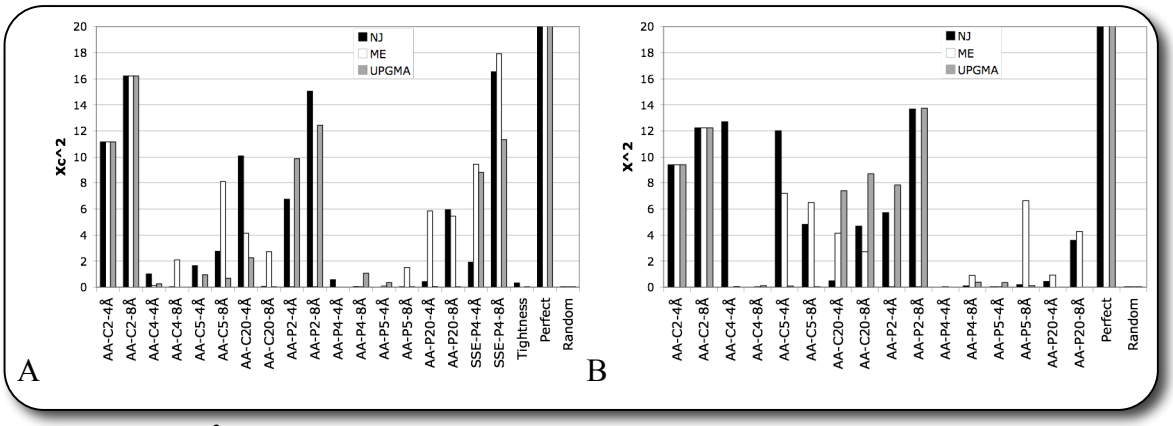


Figure 39: *χ_c^2-results for (A) 19 interface properties and (B)16 area-normalized interface properties. NJ = Neighbor Joining, ME = Minimum Evolution, UPGMA = Unweighted Pair Group Method with Arithmetic mean. AA = amino acid, SSE = secondary structure element, C2/C4/C5/C20 = composition and number of elements, P2/P4/P5/P20 = pairing propensities and number of elements, 4Å/8Å = distance-cutoff. Perfect = optimal separation, Random = random separation.*

A	H-bond forming – H-bond forming pairs	H-bond forming – H-bond non-forming pairs	H-bond non-forming – H-bond non-forming pairs
Obligate complexes	-0.059	0.276	0.057
Non-obligate complexes	0.065	0.269	-0.119

B	H-bond forming residues	H-bond non-forming residues
Obligate complexes	-0.029	0.023
Non-obligate complexes	0.024	-0.04

Table 13: *Average scores (logarithm of the counted/expected rate) of all obligate and non-obligate complexes for (A) pairing propensities and (B) compositions at a distance cutoff of 8Å.*

Surprisingly, the best score was found for a feature that was not mentioned before in any classification approach. $\chi_c^2 = 17.92$ ($\chi_c^2 = 16.52$ for NJ) was obtained by evaluating the secondary structure element-pairing propensities at a distance-cutoff of 8Å. In obligate complexes the tightly packed secondary structure element-pairs [98] such as sheet-sheet, coil-coil and, in particular, helix-helix are stronger represented than in non-obligate complexes as shown in figure 40. This leads to the tree shown in figure 41, where a circle tree is calculated and drawn by MEGA3.1 using the NJ algorithm based on distances derived from the secondary structure element-pairing propensities at 8Å. The two branches are separated after the last pairing. Counting the labels "1" and "2" as they stand for obligate or non-obligate complexes, a distribution of 18 obligate and 37 non-obligate complexes for the left branch and 57 obligate and 25 non-obligate complexes for the right branch is obtained. This leads to $\chi_c^2 = 16.52$.

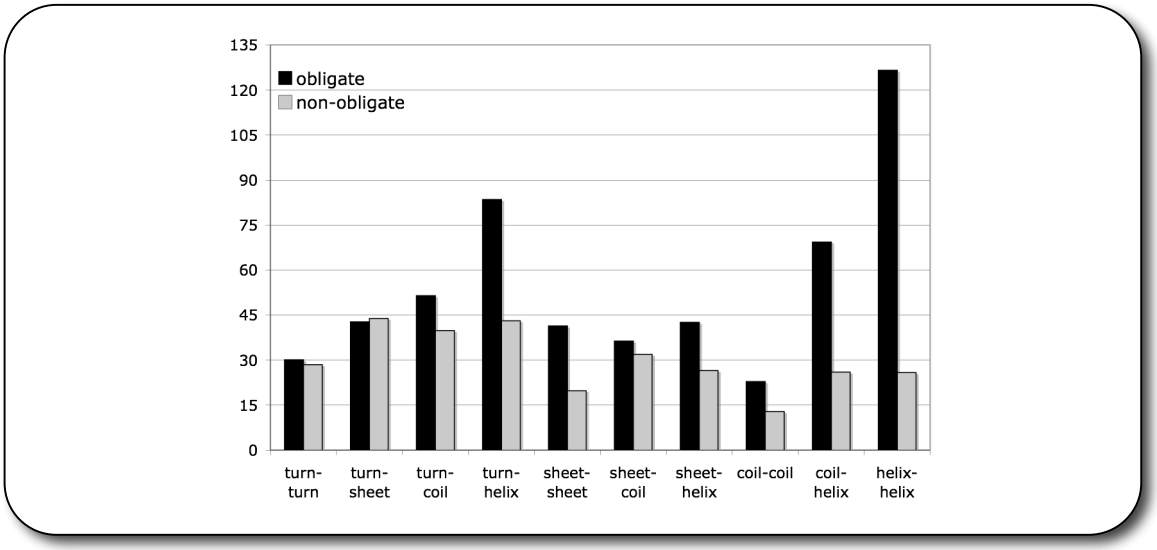


Figure 40: *The average number of secondary structure element-pairs for obligate and non-obligate complexes.*

The remaining entries in figure 39 show that the 8Å distance-cutoff for the interface residues leads in most cases to higher χ_c^2 values and thus to better obligate and non-obligate distinction than the 4Å cutoff. A general rule for the residue class is not found except for the clearer distinctions of obligate and non-obligate for the H-bond forming assemblies compared to 20 residue non-assembled data.

Independent from the clustering algorithm, the tightness of the interface defined here cannot distinguish between obligate and non-obligate complexes. Additionally, area-normalized data was used to generate distance matrices (figure 39B). Overall, the previous highly significant interface properties still scored best when area-normalization was applied. However, most χ_c^2 values dropped by approximately 20%. In three cases – for the residue composition at 4Å distance cutoff and residue classes 4 and 5 – the χ_c^2 value shows a clear increase when the data was area-normalized.

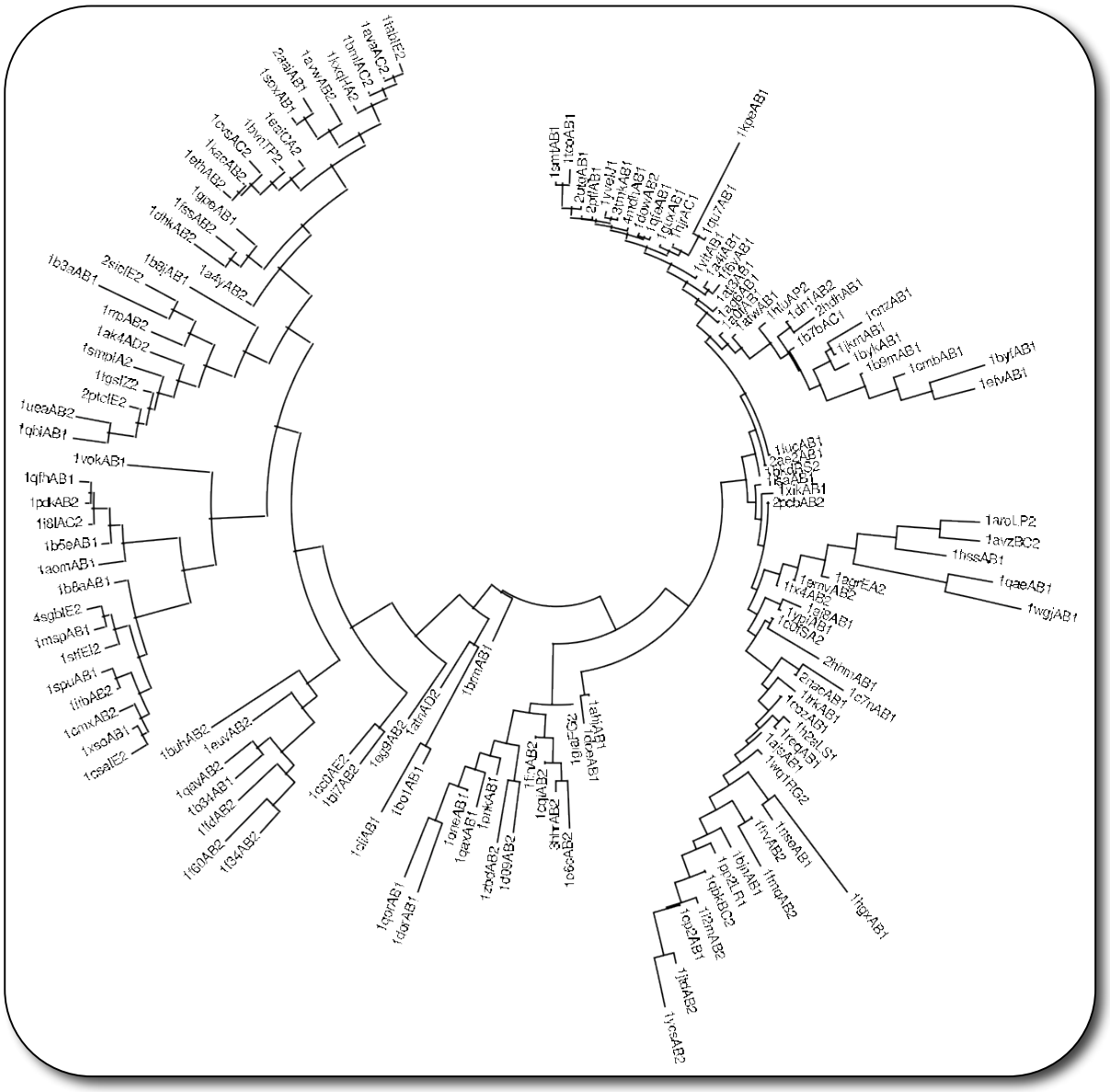


Figure 41: *Circle tree drawn using MEGA3.1 based on the NJ clustering algorithm and secondary structure element-pairing preferences at 8Å distance-cutoff. Code: <pdbcode><chaincombination> <complextype> 1 = obligate; 2 = non-obligate.*

4.3.2. Evaluating the Clustering Algorithms

Distances between the complexes for given properties were calculated and used for clustering. After the first bifurcation the distributions within each cluster were divided into two groups and statistically evaluated. 19 interface properties were analyzed. This section analyzes the effect of the three different clustering algorithms on the separation of the two complex types. As figure 42 shows, the correlation of the calculated distances for

secondary structure element-pairing propensities at a distance cutoff of 4Å and 8Å is very high (0.9). In other words, a very similar distribution of secondary structure element-contacts is found when considering secondary structure element-pairs in direct contact (4Å) or at more distant contacts (8Å).

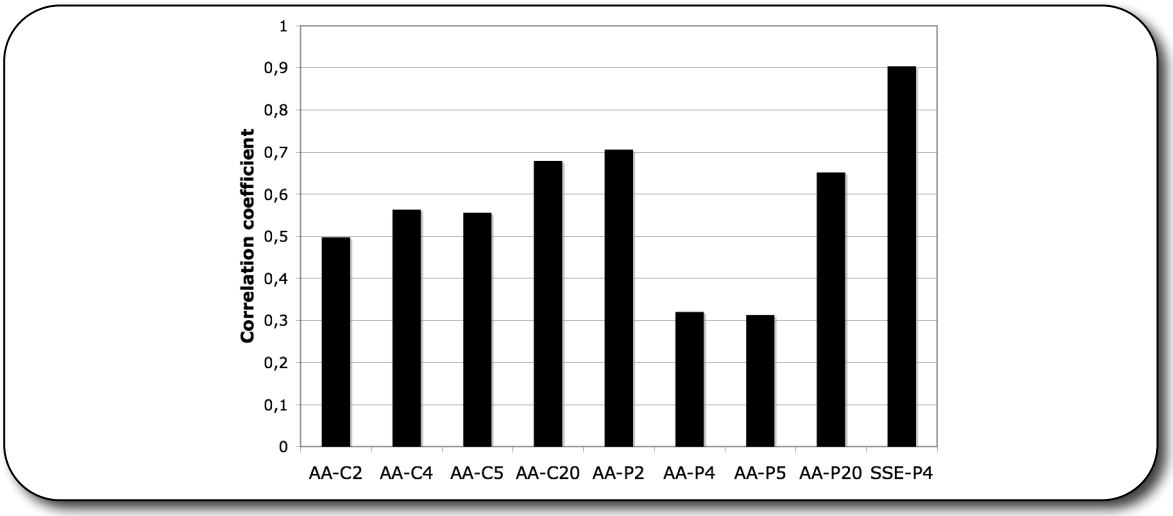


Figure 42: *Correlation coefficients of the interface properties using a distance cutoff of 4Å or 8Å. AA = amino acid; SSE = secondary structure element; C = composition; P = pairing propensities; 2, 4, 5, 20 = number of elements.*

This finding is unexpected. Figure 43 shows the relative distribution of secondary structure element-pairs at 4Å (A) and 8Å (B) distance cutoff. The distributions are, as stated in figure 42, nearly the same.

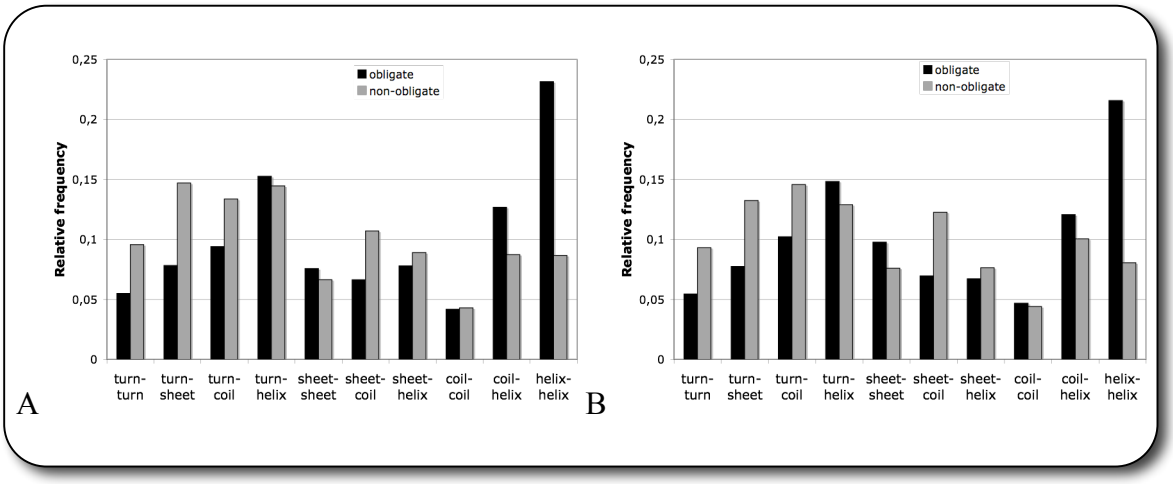


Figure 43: *Relative distribution of secondary structure element pairs for obligate and non-obligate complexes at a distance cutoff of (A) 4Å and (B) 8Å.*

However, in spite of this high correlation, χ_c^2 equals 1.91 for 4Å and 16.53 for 8Å distance cutoff (figure 39A). This is another unexpected observation, as the almost identical distance pairs should lead to the same distribution of the clustering for a given algorithm. Therefore, the relation was tested between the correlation coefficients of 4Å and 8Å data and the $\Delta\chi_c^2$ at 4Å and 8Å distance cutoff for each clustering algorithm. The results are shown in figure 44. Apparently, the NJ algorithm is very sensitive toward the clustered data. The more similar the clustered data is, the larger is the $\Delta\chi_c^2$ value. This tendency is less pronounced for the ME and smallest for UPGMA algorithm. The normalized relation in figure 45 shows the same trend. One may suspect this to be a peculiarity of the χ_c^2 test where small changes in the data separations may yield large χ_c^2 changes. Figure 46 shows how the χ_c^2 values change once the separation in two categories is systematically changed from one extreme to the other. As expected, the χ_c^2 test is very strict in highly significant areas. In the current case, a χ_c^2 change from 16.53 to 1.91 can be caused with approximately 8 displacements. This suggests noticeable differences in the separations which derive from the NJ algorithm using the secondary structure element-pairing propensities at a distance cutoff of 4Å and 8Å. Table 14 shows the actual distributions, where the almost identical pair distance-lists lead to clearly different clusters after the first branching point.

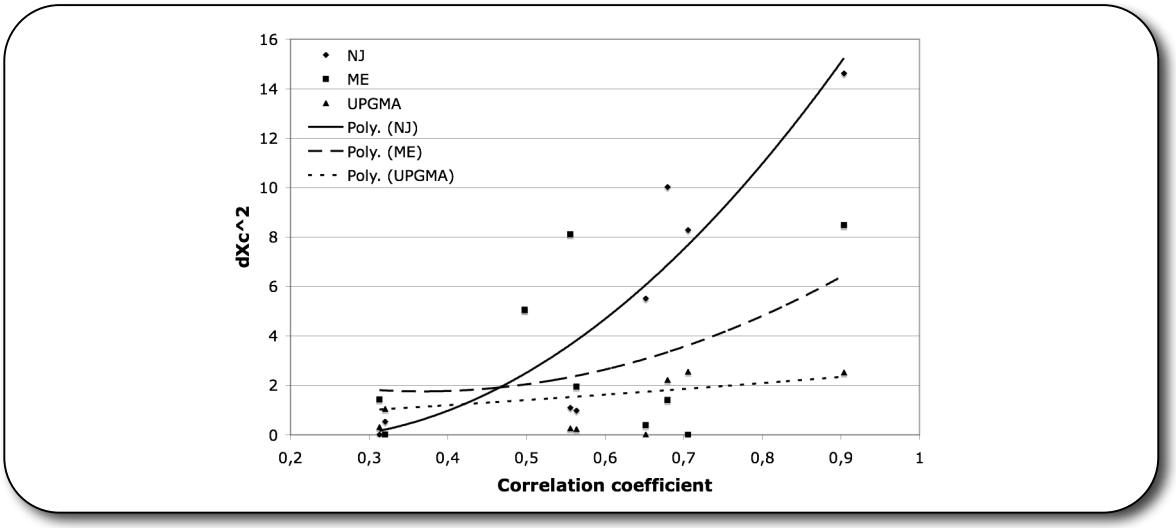


Figure 44: *Correlation coefficients for various interface properties computed using either 4Å or 8Å distance cutoffs vs. the absolute difference of the corresponding χ_c^2 values, termed $\Delta\chi_c^2$.*

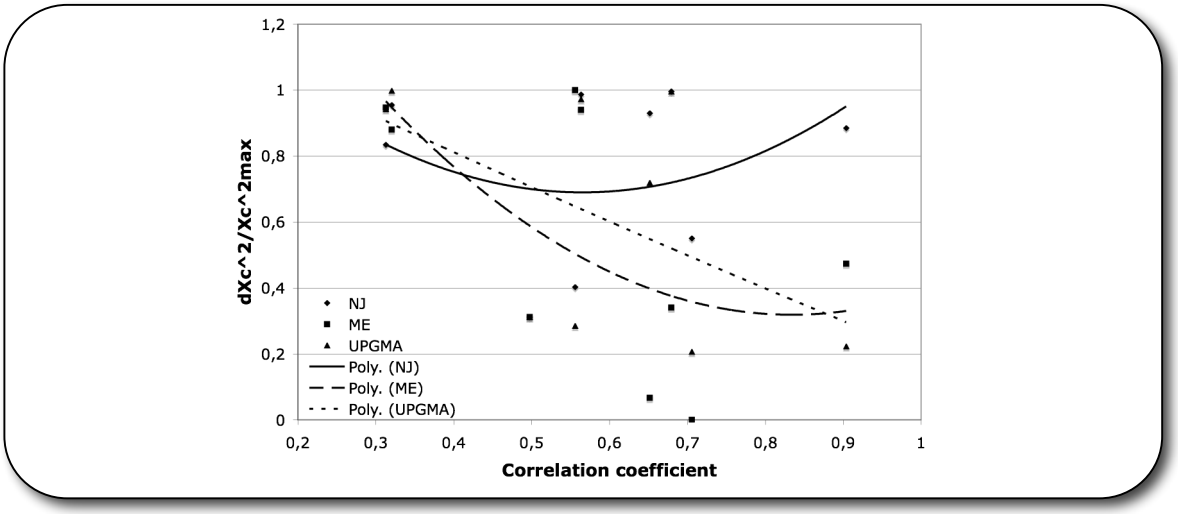


Figure 45: *Correlation coefficients for various interface properties computed using either 4Å or 8Å distance cutoffs vs. the relative difference of the corresponding χ_c^2 values, where $\Delta\chi_c^2$ is divided by the larger χ_c^2 value for the two distance criteria.*

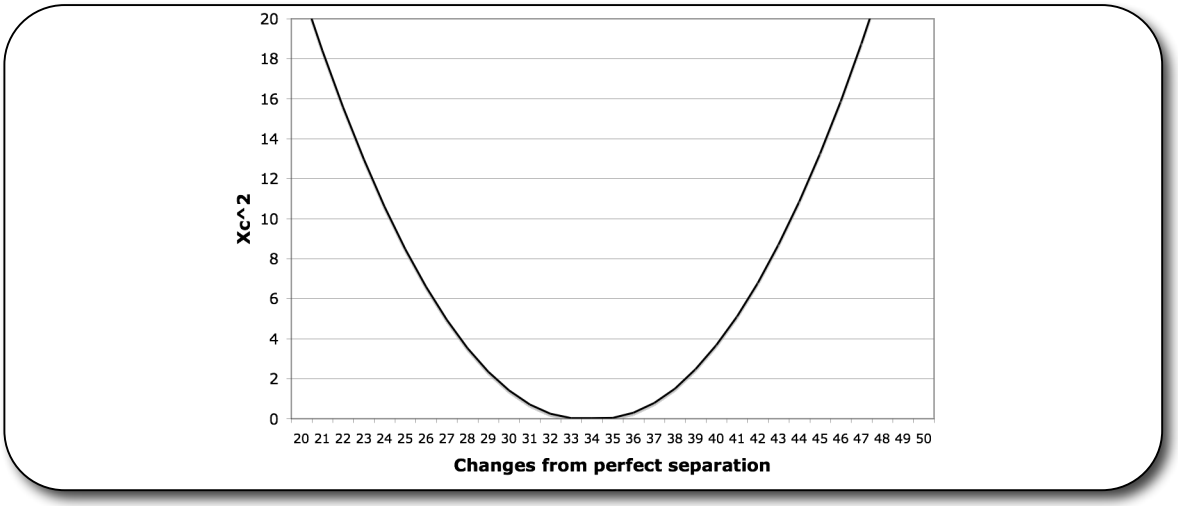


Figure 46: *Change of χ_c^2 values as the separation in two categories is systematically changed from one extreme to the other. An x-axis value of X indicates a separation of 75-X, X, X, 62-X (e.g. 20: n_{1obl}=55, n_{1non}=20, n_{2obl}=20, n_{2non}=42).*

	SSE-P4 4Å	SSE-P4 8Å
Group 1 obligate	23	18
Group 1 non-obligate	27	37
Group 2 obligate	52	57
Group 2 non-obligate	35	25

Table 14: *Clustering distribution after the first branching point for NJ using secondary structure element-pairing preferences at 4Å and 8Å distance cutoffs.*

4.4. Conclusion and Outlook

Applying three common clustering algorithms in order to characterize protein-protein interactions based on 19 different interface properties, a set of 75 obligate and 62 non-obligate complexes was clustered to identify interface properties with a high sensitivity to the distinction of obligate and non-obligate interactions. Interface properties leading to clear separations of the two interaction types will also lead to conserved patterns within the interface regions.

Residue compositions and pairing propensities for different residue assemblies together with secondary structure element-composition and pairing propensities, and the tightness of the interaction were used to compute distance matrices among all structures in the dataset. The distance matrices were used for clustering by applying three different algorithms: NJ, ME and UPGMA. Evaluating the reliability of the clustering methods allows concluding that the consensus of alternative clustering methods provides a more balanced basis than individual approaches. However, due to the focus of the clustering on interface area only, the criterion defining this area plays an even more important role than the clustering method. For that, two different distance cutoffs between the heavy atoms of the interacting chains were used. A distance cutoff of 4Å mainly considers tightly bound residues and regions. Water penetrations are discriminated. Using a distance cutoff of 8Å will also include contacts mediated by interfacial water molecules and include peripheral electrostatic interactions. The results suggest that averaged over all three clustering algorithms the distance criterion of 8Å leads to a better distinction of obligate and non-obligate complexes. However, this might as well have another explanation; the number of interacting residues at 4Å distance cutoff is rather low for the composition of all residues or secondary structure elements and even lower for their pairing propensities. Based on these results it can therefore not clearly be concluded, whether the interface representation is better defined with an 8Å distance cutoff or not. Due to the more significant number of counted interaction pairs the results of the χ_c^2 tests based on the distance cutoff of 8Å are summarized in the following way:

At a level of at least 99.9% confidence three interface properties result in a significant distinction of obligate and non-obligate complexes. It should be noted that this confidence interval has nothing to do with accuracies of predictions for classification

approaches. The residue-pairing propensities score in a χ_c^2 value of 15.03 (13.72 when area-normalized data was used) in the case of pairing propensities of H-bond forming residues and those that do not form such bonds. Taking a closer look at the average properties reveals the high importance of H-bond forming and non-forming residues at interfaces. Obligate interactions mostly consist of H-bond non-forming interactions while non-obligate interactions mostly form H-bonds. This finding is closely related to the next highly significant distinction criterion: composition of H-bond forming and non-forming residues. A χ_c^2 value of 16.2 (12.22 when area-normalized data was used) was obtained. This is in agreement with the current opinion on the importance of hydrophobicity at the interface of protein-protein interactions. Obligate protein-protein interactions must be stickier and require a higher hydrophobicity of the interface. However, in this approach the hydrophobicity of protein-protein interfaces did not lead to the best distinction of obligate and non-obligate complexes. In chapter 2 the distribution of the secondary structure elements did not show clear results, where the secondary structure element-pairing propensities led to propensities following the rules of tight packing. In this chapter, however, the use of secondary structure element-pairing propensities scored in a maximum χ_c^2 value of 17.92, which is the highest significance found in these analyses. Although steric complementarity plays an important role in non-obligate interactions as well, a stronger proportion of tightly bound secondary structure element-pairs for obligate complexes was observed. This supports the stickiness of obligate complexes and meets our expectations. The additional criterion for the tightness of fit based on the number of interface residues at a distance cutoff of 4Å and 8Å led to a completely random distinction as the χ_c^2 value scored in 0.31. Apparently, this property is a too simplified model for the tightness of an interaction.

Although none of the employed interface properties have led to a desired clear separation and therefore to a conserved pattern within the interface region of different complex types, the important role of interface hydrophobicity and tight packing at the interface area was underlined. Additionally, a new method was applied to visualize clusters by dendrograms. This is an easy, yet clear way to visualize even complicated clusters. It can also be used to graphically show the distribution of properties with dimensions higher than three such as residue composition and pairing propensities. Many current support

vector machine-approaches lack visualization techniques for more than three features [129][130][131][62]. Such dendrograms allow almost unlimited dimensionality for the feature vectors since they are based on correlation coefficients. Furthermore, the construction of such trees provides the opportunity to determine new classifiers. Depending on the level of bifurcation and penetration of the tree, one could determine different types of complexes with a different rate of interface property-purity.

In chapter 5, the dataset will be increased in order to increase the statistical strength. Based on a larger dataset and efficient classification approaches, such as support vector machines, a larger number of interface properties shall be analyzed and combined. This may finally lead to a clearly conserved pattern for transient interfaces by efficiently separating transient from obligate complexes. Given such clear patterns, the docking problem may be addressed in a new attempt.

5 A Database for Analyzing Biomolecular Contacts

5.1. Overview

Databases are used to store information in a systematical way and allow quick and flexible access to the data. In this chapter a large number of information based on protein-protein interactions is retrieved and systematically stored in a MySQL database-server. Mainly transient/non-obligate and permanent/obligate complexes were collected from the literature and a tcl/tk script was employed to retrieve a large number of information. The database currently contains 534 interfaces extracted from 479 PDB files, where nearly half of the interfaces are from transient/non-obligate and the rest from permanent/obligate complexes. The database will facilitate further more detailed statistical analyses in order to find clear patterns in complex types.

This project was tackled in cooperation with Peter Walter during his diploma thesis and current PHD thesis. Peter Walters' work was to store the data into the MySQL database and modify the structure of the database for the extended data. Furthermore, he constructed a user-friendly interface for public use.

5.1.1. Introduction

Currently, there is a strong need for methods that would help obtaining an accurate description of protein interfaces in order to be able to understand the principles that govern molecular recognition and protein function. While many of the recent efforts are focused on computationally identifying and characterizing protein networks and need to extract information on protein interaction from the PDB, these data are quite hard to access directly from the PDB database. Therefore, a number of groups have developed databases that store different aspects of protein-protein complexes. The group of Michael

Schröder in Dresden for instance has developed the SCOPPI (Structural Classification Of Protein-Protein Interactions) database. It was published recently and contains interactions between protein domains [132]. These domain interactions are derived from all known protein structures and are classified and annotated. Applying a distance criterion retrieves inter-domain interfaces. Furthermore, their database contains various interface characteristics such as number, type and position of interacting amino acids, conservation, interface size, and permanent or transient nature of the interaction. Another interesting database from the group of Mayte Pisabarro at the same institute was published in 2006 as well: SCOWLP (Structural Characterization Of Water, Ligands, and Proteins) [133]. This database is developed for characterization and visualization of the PDB protein interfaces and includes proteins, peptidic-ligands, and interface water molecules, as descriptors of protein interfaces. The web-server allows structural analysis and comparisons of protein interfaces at atomic level. SCOWLP is automatically updated with every SCOP release. However, earlier approaches such as the Biomolecular Interaction Network Database (BIND) from Bader et al. are much wider spread [134]. This database is designed to store full descriptions of interactions, molecular complexes and pathways. Additionally, chemical reactions, photochemical activation and conformational changes can be described. Everything from small molecule biochemistry to signal transduction is abstracted so that graph theory methods may be applied for data mining.

Here, a new database is presented, which is based on previous analysis and results of this thesis. Modifying and extending the tcl/tk script from chapter 2 and considering up to 7 different interface criteria led to a large number of interface properties. The data is stored in a MySQL database, mainly to quickly access detailed information on protein-protein interaction sites and also perform additional calculations for statistical analyses and extended output functions. Furthermore, the database is extended with a user-friendly web-interface, which allows public access to the data.

5.2. Structure

5.2.1. Data Set

Six sources for obligate/permanent and non-obligate/transient complexes were taken from the literature [135][136][128][127][67][69]. Overlapping data was deleted and few contradictory classifications manually corrected. After applying quality filters on the dataset, such as unique residue labels for given sequence positions and proper atom labels, a set of 534 structures was retrieved.

5.2.1.1. Protein-Protein Interaction Data Retrieval

Based on the molecular visualization program VMD a previously employed script (see chapter 2) was extensively modified. After loading the PDB files, the script examines in its first section all residues of each chain and generates a list consisting of each sequence position, secondary structure element-descriptor, and the residue descriptor. Furthermore, the script analyzes whether the residue lies on the surface, core, or interface region of the protein for a given criteria. Calculating the accessible surface-area contributions of each residue based on a probe with radii of 1.4Å or 4.0Å (see section 4.2.1 and table 12), those with contributions larger than $0Å^2$ are understood to lie on the protein surface. As a probe size of 1.4Å yields larger surface area-contributions for nearly all residues, the probe size of 4.0Å is used for determining surface residues. Additionally, the change of accessible surface-area upon complex formation is calculated for each residue. When the accessible surface-area changes upon complex formation, the residue is said to be involved in the interface region. This generated list is also used to store the sequence of the protein chain in the database in order to allow more analyses based on the protein sequences as well. Furthermore, the interface size is calculated in $Å^2$ based on the buried accessible surface-area upon complex formation. Both probe sizes are considered. In the next section of the script, distances between pairs of atoms are measured and a number of distance cutoffs are used to retrieve a list of atom pairs. For these distance-based criteria, thresholds such as 4Å, 5Å, 6Å, 7Å and 8Å are used. In the previous chapter the distance cutoffs 4Å and

8Å were employed. While 4Å considers only very tightly bound regions of the interface, barely allowing water penetration, a distance-range of 8Å even includes peripheral electrostatic interactions. Here, the distances in-between these two extremes are evaluated as well.

When using larger distance-cutoffs, many buried residues that most likely do not participate in the interaction are considered as well. Such interactions are discriminated in the same way as described in the previous chapter. All gathered atom pairs are recorded together with their chain identifier, residue descriptor, residue position in the sequence, and secondary structure element-descriptor. Storing the data at atomic level also allows retrieval of information on side-chain and backbone interactions.

Considering the interface criteria from the first section of the tcl/tk script, 7 different criteria for interface residues are analyzed. Specifying an XML format for importing the data into the database, the script outputs all collected data in a properly organized XML file.

5.2.1.2. Additional Data

When storing computationally retrieved information on protein-protein interactions, experimental data may be very valuable to combine with. Kinetic data on the available interactions clearly determines the strength of interactions and may therefore be helpful in finding the best interface criterion or more clearly distinguish permanent and transient complexes. Since available kinetic data is rather limited and difficult to retrieve, such data is available only for a small number of complexes in the database. Carla Haid collected this information. However, it turned out that most experimental kinetic values have been measured at quite different experimental conditions, such as pH value, pressure and temperature, which limit their comparability.

In addition to kinetics data, CATH, SCOP and UniParc identifiers were also added to the corresponding entries in the database. Such cross-references to the popular structural classification databases and the largest protein sequence database may facilitate more extensive analyses.

5.2.2. Database Design

Currently, the database includes a number of relations that do not yet fully store all gathered data from the tcl/tk script. The central relation is ‘datasets’ (see figure 47). Every entry in ‘datasets’ consists of one interface specified by the PDB identifier, the combination of chains, and the size of the interface after a number of different interface criteria. The ‘contacts’ relation contains the list of interacting residue pairs for the five given distance-based interface criteria together with the corresponding secondary structure element. For each residue pair and the given distance-based interface criterion the side-chain and backbone-pairing propensities are stored in ‘sidechain/backbone’. The ‘sequence’ relation stores the entire sequence and for each position in the sequence, the amino acid and secondary structure element-descriptor, the accessible surface area of the residue after a probe size with 1.4Å and 4.0Å radius before and after complex formation. Based on the five distance-based interface criteria, the relations ‘aa composition’, ‘sidechain/backbone composition’, and ‘sec. struc. compostion’ store the composition data on residues, side-chain backbone, and secondary structure elements for each chain and result in 10 entries for each interface that is stored in ‘datasets’. The ‘SCOP’, ‘CATH’, and ‘UniParc’ relations contain the identifiers for the corresponding databases. As SCOP and CATH identify domains instead of chains, there is a relation of M:N to ‘datasets’.

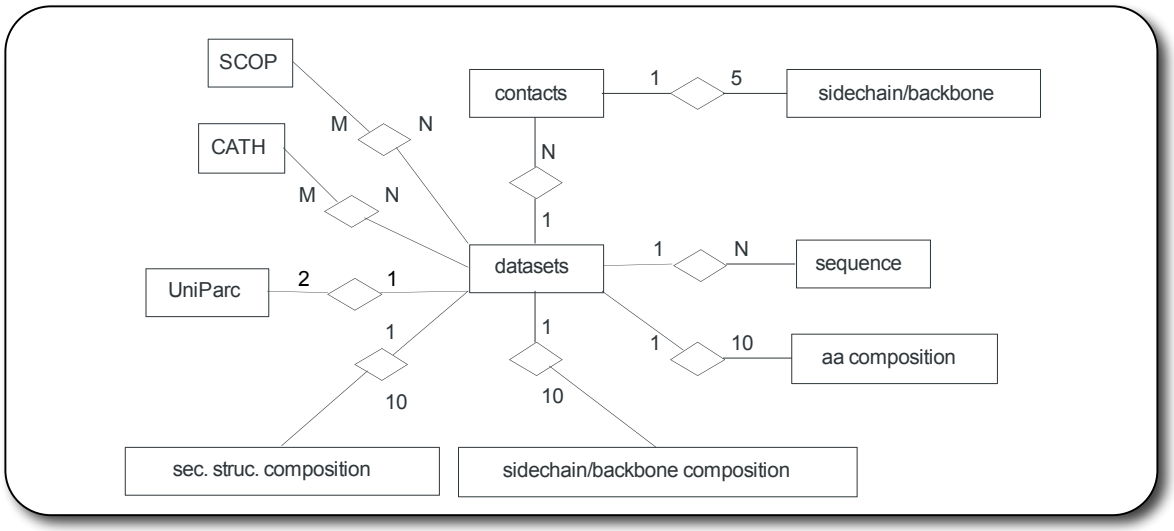


Figure 47: *Dataset structure. 10 relations are shown containing different types of data. 'datasets' is the central entity in the schema and represents the interfaces. Every interface may be assigned to more than one CATH, SCOP and exactly two UniParc classification. A CATH, and SCOP classification may refer to one or many interfaces (N:M). Each interface in 'datasets' consists of a number of amino acid-pairs that are represented by the 'contacts' (1:N). A contact pair contains a number of side-chain and backbone compositions depending on the distance criterion. 5 distance criteria are considered and lead to a 1:5 relation. Similarly, the relations between 'datasets' and 'aa composition', 'sec. struc. composition', and 'sidechain/backbone composition' result in a 1:10 relation as they refer to each of the two chains. An interface with its chains consists of a number of residues leading to the 1:N relation between 'datasets' and 'sequence'.*

5.2.3. Database Administration

Based on the XML files from the tcl/tk script an import filter was developed under JDOM, a class extension that offers extended functionality in importing and exporting XML-based data. Additional data such as the CATH, SCOP, and UniParc identifiers were added semi-automatically. Downloading the parseable datasets of CATH, SCOP, and UniParc, simple parser were prepared to output a table of these identifiers for given PDB files and chain identifiers in the current dataset. This table was imported in the database. The same procedure is followed for kinetic data where information about the experimental conditions was also stored. The developed administrators web-interface based on java allows the modification, deletion, and insertion, as well as the detailed query of the data.

5.3. Features

5.3.1. Query Options

The administrator interface allows detailed and individual queries. A number of standard queries were prepared for public use by simply specifying the query content and the area of searching. Typically, the user may search for specific PDB entries by their identifier, description, and links to the additional databases such as CATH, SCOP, and UniParc. However, it also facilitates to select interface properties and retrieve the complexes suiting the specified criteria. The user may define the interface criterion and the property of the interface such as hydrophobicity, rate of charged residues, size, rate of certain amino acids or secondary structure elements, values for kinetic data, side-chain backbone rates, and more. Allowing combinations of queries with logic operators, the output of the query can become very specific to the interest of the user. The user further has the option to refine the query results by re-applying all previously specified filter options. Figure 48 shows a screenshot of the current query options.

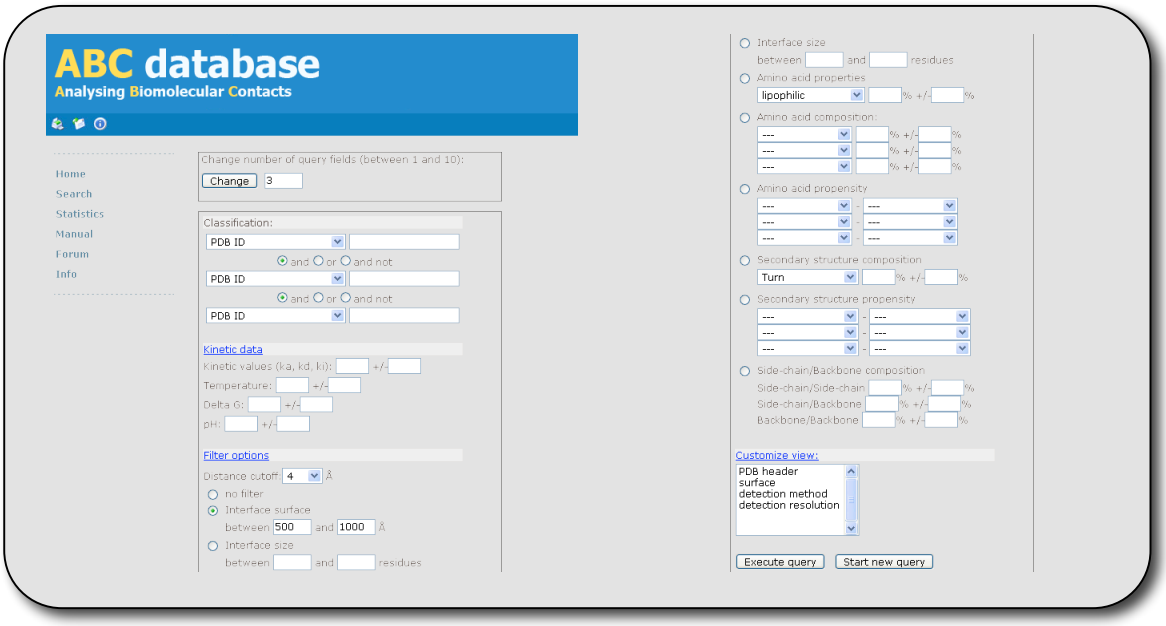


Figure 48: *Query options.*

5.3.2. Data View

For a specified query, the user will retrieve a list of PDB identifiers with the particular chain combination for the interface. Furthermore, the PDB description that is stored in the *HEADER* compartment of the PDB file will be listed, as well as links to the CATH, SCOP, UniParc, and directly to the RCSB PDB entries that are related to the displayed complex (figure 49). Information mainly originating from the relation 'dataset' may be displayed as the users can specify the display options (figure 48). However, a number of second-level display options are offered too. More detailed information can be viewed by simply clicking on a listed entry (figure 50). For the specified interface criterion all interface residues stored are projected to the protein sequence of each chain. A number of statistical analyses may be viewed in figures generated by JfreeChart based on JAVA. Such analyses show the residue composition of the interface for the specified interface criterion, as well as the frequencies of hydrophobic, hydrophilic uncharged, and charged residues. The three-dimensional structure of the protein and the interface may be viewed using the Jmol-applet. This applet is combined with information on the interface area that can be projected on the displayed protein structure.

One of the most interesting features of the database is the possibility to also display the residue and secondary structure element-pairing propensities in a color matrix, where different fraction methods can be calculated and interface criteria changed. The next section shows more about this feature.

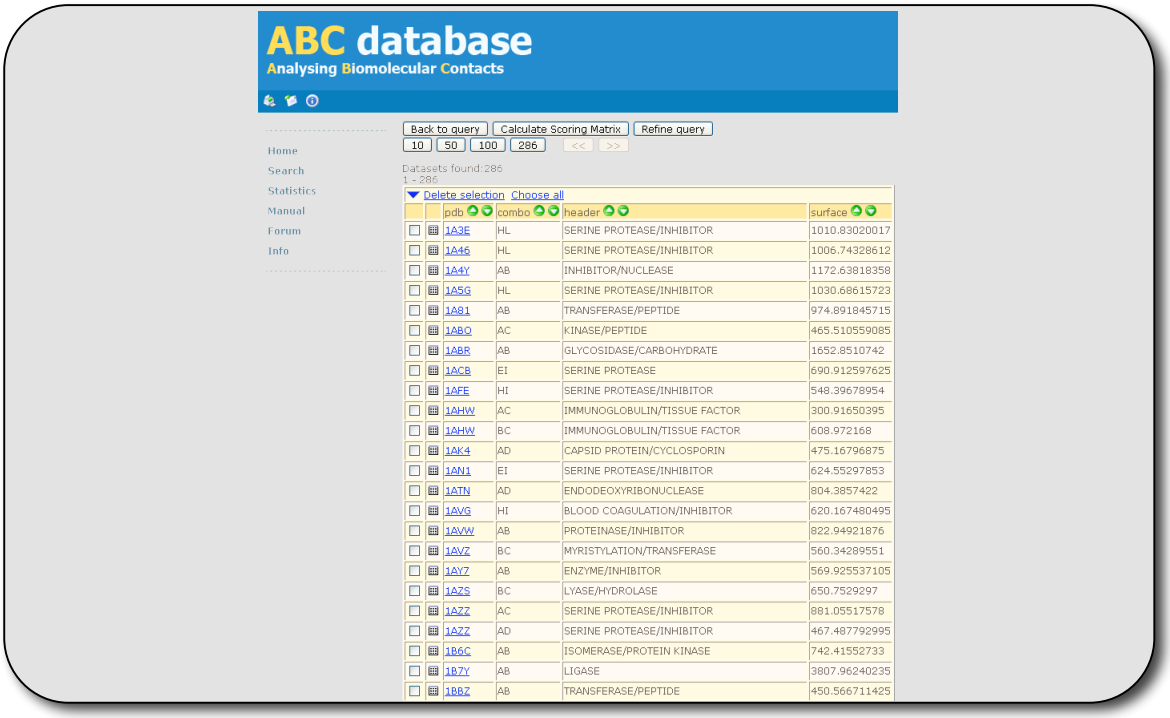


Figure 49: *Output of the results for a given query parameter.*

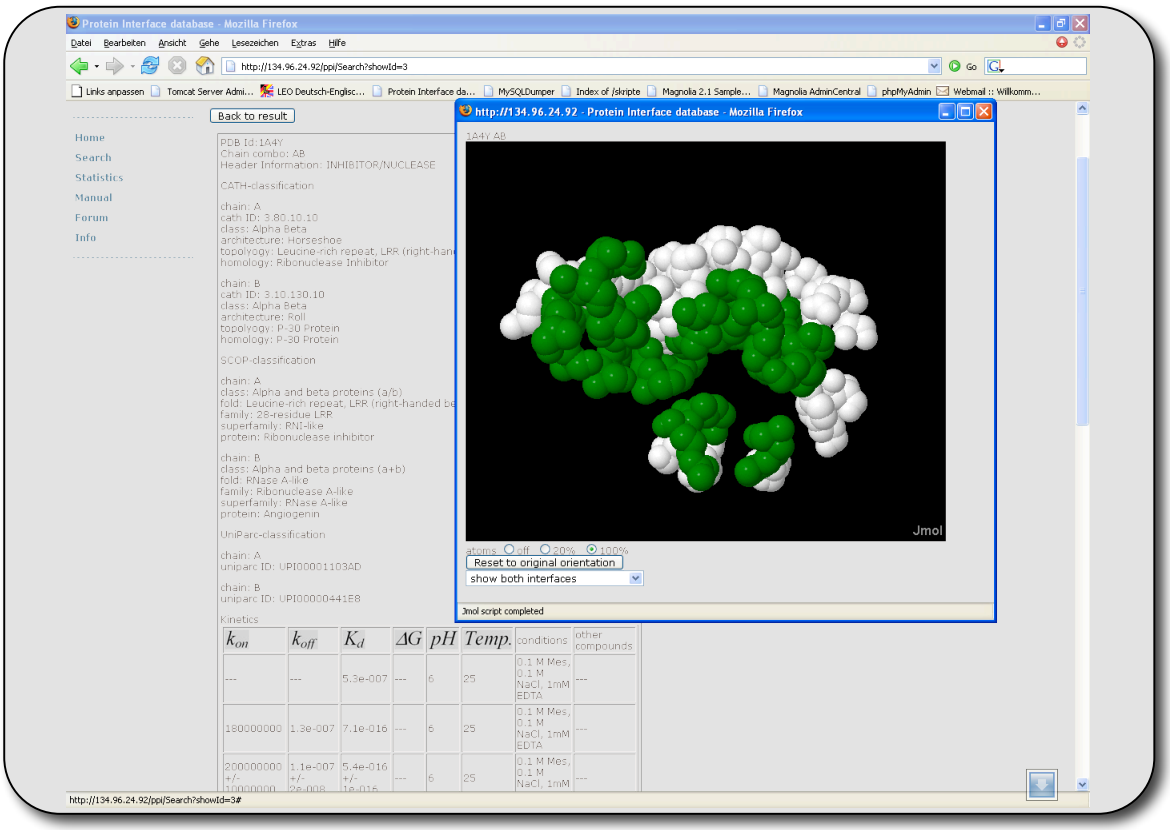


Figure 50: *Detailed result output for a given interface with Jmol protein and interface view.*

5.3.3. Output Options

As mentioned in the previous section, an option to view residue and secondary structure element-pairing propensities was included. This feature serves the purpose to quickly compute scoring matrices as they were used in chapter 3 for the protein-protein docking approach. Based on a selected list of complexes, a given interface criterion, and fraction method, pair potentials may be computed in the same manor as introduced in section 3.2.2.2. This function allows quick generation of scoring matrices based on varying datasets and therefore facilitates testing the relation between specificity of the dataset and its predictability on suitable benchmark sets. Currently supported output formats include plain text or Microsoft Excel format.

5.4. Outlook

Although the database is not yet ready for public use, most of the interaction data is stored properly and may be queried by the administrative interface. This will enable exhaustive analysis on a large dataset and retrieval of rich data on interface properties in order to eventually find conserved patterns within the two types of complexes. Such conserved patterns may not only simplify the expansion of the dataset in a fully automatic way, but also allow computing scoring matrices based on characteristic interface properties to support the predictability of docking approaches such as BDOCK.

The next step will be to include atomistic interface pairs, as they were collected. Atomic contacts were used in a number of prediction approaches and seemingly lead to significant results [69][137]. Also, some important features that have previously shown an increase in significance of the analysis have not yet been integrated, such as the residue classes proposed in table 12. Compulsory extensions are the implementation of BioJAVA to increase the analysis strength of the data. BioJAVA contains a number sequence-alignment functions that may be used for internal computation of conservation scores for the residues, as well as allow the discrimination of data redundancy, as this has not yet been addressed in the dataset. Also the limited number of complexes that are

covered with kinetic data will be increased and the retrieval of links to other database should be automated.

Beside the non-obligate/permanent and obligate/transient complexes collected so far, other more differentiated complex types may be introduced, as well as the popular differentiation of interfaces into rather “wet” or “dry” contact area [138]. Other types of biological contact elements such as interactions between proteins and small molecules or protein and DNA may become another future focus in this database project.

6 Classification of Obligate and Non-obligate Complexes

6.1. Overview

A powerful database presented in chapter 5 currently contains 534 obligate/permanent and non-obligate/transient complexes. Storing interaction data from 7 different interface criteria meets the requirements for more exhaustive clustering and classification approaches. In this chapter the dataset from the previous chapter is used to compute up to 347 interface properties (feature vectors) in order to find the best descriptor for separating obligate/permanent and non-obligate/transient protein-protein interactions. A support vector machine is trained to identify the interface property with the highest significance in separation of the two types of complexes. This machine learning project is based on the experiences gained in chapter 4, and provides a significant extension. Here, a larger dataset, more interface properties, and a more efficient classification is applied in order to find combinations of properties leading to highly significant separations of the data. The results of this work will on the one hand ease the further expansion of the current data in the database, and on the other hand reveal properties or property combinations that lead to strongly conserved patterns within the interface region of the complexes, as it was also addressed in chapter 4.

6.1.1. Introduction

In recent years a lot of efforts were dedicated to the investigation of protein-protein interactions. Although there are many approaches identifying the general physicochemical properties of protein complexes [24][17][16][139][140][72][106][141][142], little is yet accepted as common knowledge. Other studies showed that protein-protein interactions might be highly specific and diverse at the same time [93][16][143]. Therefore, modulating the interactions has become of great interest. A fundamental distinction of

protein-protein interfaces is the separation into obligate/permanent and non-obligate/transient complexes. Although the general properties of these complex types seem clearly different, biological systems, such as protein complexes, do not obey in all cases a straightforward classification. The first attempt to classify protein-protein complexes into obligate and non-obligate interactions was done by Mintseris et al. in 2003 [69]. The authors introduced the concept of atomic contact vectors. Compiling a dataset of 345 structures comprising 147 obligate and 198 non-obligate complexes, a prediction accuracy of 91% was achieved. In 2006 Zhu et al. introduced NOXClass that is a web-service for predicting protein-protein interaction types [68]. In their work a dataset of 75 obligate and 62 non-obligate structures was used to retrieve six properties, which were combined using a support vector machine approach [144]. By combining features such as the interface area size, relative interface area size and the area normalized residue composition, the authors obtained an accuracy of 88.32% for the separation of obligate and non-obligate complexes. However, this work mainly focused on the separation of obligate, non-obligate, and crystal packing structures where an accuracy of even 91.8% was achieved. In a very recent work, Block et al. achieved an accuracy of 93.6% for classifying permanent and transient complexes using C4.5 decision trees on a data set of 147 permanent and 198 transient complexes based on the compiled list of Mintseris et al. [69]. The authors calculated two different atomic contact vectors, DrugScore pair potential vectors and SFCscore descriptor vectors and used four different machine learning algorithms: SVM, C4.5 Decision Trees, K Nearest Neighbors, and Naïve Bayes algorithm [137]. Three different feature selection methods were used to quickly find the best combination of feature vectors and achieve the highest accuracy. Similar to the results of Mintseris, the atomic contact vectors led to the best separation. Another very recent study on classifying permanent and transient protein interactions was published by Kottha and Schröder [145]. These authors used a dataset of 161 permanent and 242 transient interactions and calculated more than 300 interface attributes (features) mostly related to size, physicochemical properties, interaction propensities, and secondary structure elements. A prediction accuracy of 97% was achieved by applying support vector machines to the molecular weight difference of the interacting chains, size of the buried surface and number of hydrophobic contacts. The molecular weight

difference alone resulted in an accuracy of 80%.

In summary, the previous studies did not identify a unique property being able to separate obligate and non-obligate complexes. However, highly significant distinctions can be achieved by combing a small number of features and applying machine learning algorithms such as support vector machines.

In this work, 347 feature vectors containing 9,692 features are computed and a non-redundant dataset containing 251 obligate and 212 non-obligate protein-protein complexes is used for training and testing. The feature vectors contain a large number of residue compositions and pairing propensities, where the interface area was defined by different criteria. Furthermore, different normalization methods were applied. The final goal is to establish an effective filter for the separation of obligate and non-obligate protein-protein interactions, which will be used in upcoming projects in order to classify the entire content of the RCSB PDB to retrieve more complexes for further studies. The results will also address the previous attempts to properly define transient complexes and enhance the sensitivity of rigid-body docking approaches.

6.2. Methods

6.2.1. Dataset

534 structures containing obligate/permanent and non-obligate/transient complexes were previously collected and stored in a database (see chapter 5). Possibly, this database may contain protein-protein complexes of identical or highly similar sequences. For the analyses in this chapter, such redundancies may mislead the results and were therefore excluded. A common method for defining data redundancy is to apply a sequence identity threshold of, for example, 25%. Sequences with higher levels of identity are thought to be homologs and thus excluded from the analysis. As this project focuses on interface areas, redundant data is defined as such that have correlation coefficients among their residue pairing propensities of more than 0.8. Preliminary tests in dealing with correlation coefficients have shown that even at values up to 0.75, random correlations are still possible, although the probability is very low. Due to this observation and the urge for a

large dataset, the balance of possible redundancy and large dataset was set to a correlation coefficient threshold of 0.8. Also, it was found that within complexes with correlation coefficients beyond 0.8, no relation is found between the correlation coefficient and the level of sequence identity. This indicates that similar structures do not necessarily form the same interface and vice versa. After removing redundant data the size of the dataset equals 463 complexes containing 251 obligate/permanent and 212 non-obligate/transient structures. This dataset is not only large but also fairly balanced, which is a good basis for this analysis.

6.2.2. Construction of the Training and Test Set

Based on a dataset of 463 protein-protein complexes, 347 properties (feature vectors) with 9,692 features were either provided by the database or extracted by scripts. The R package e1071 [146][147] interfacing to libsvm [148] was used to perform the support vector machine classification.

6.2.2.1. Interface Criteria

Up to 7 interface criteria were defined and employed (see also chapter 5). For the distance-based criteria threshold values of 4, 5, 6, 7 and 8Å were used. To avoid buried interfaces counted as interface residues, a residue fulfilling the distance-criterion also has to have a larger surface contribution than 0Å^2 when a probe with a radius of 4Å is used to calculate the surface area (for further details see section 4.2.1). As for surface area based criteria, an interface residue was collected when its accessible surface area changed upon complex formation. The accessible surface area was computed using probes of 1.4Å or 4.0Å radius.

6.2.2.2. Fraction Methods

A common way to compare residue compositions or pairing propensities is to generate statistical potentials. In opposite to simply counting residues or pairs of residues, a statistical potential, similar to a lod score, is based on the logarithmic rate of the evaluated number and an expected number. There are different ways to define the expected number. One common procedure is based on the frequency of a residue to occur in the protein accounting for the different frequencies of amino acids (figure 18). Here, three different methods are applied to calculate the expected values. First, the mole-fraction method, is proportional to the fractional abundances of the residues or secondary structure elements or their pairs. Second, the contact-fraction method, is only applied to pairing propensities where it is proportional to the frequencies of the two residues or secondary structure elements to be involved in any pairs. The third method is the area-fraction, which is – similar to the mole-fraction – proportional to the relative area contribution of the residues to the surface area of the protein and was used in the work of Zhu et al. as well [68].

Furthermore, the mole-fraction and area-fraction methods are applied in several variations. The mole-fraction is extended by relating the fractional abundances of the residues or secondary structure elements to (a) the full protein sequence of all complexes; (b) the surface sequence and (c) the interface region of all complexes only, which is based on a probe size of 4.0Å radius. The statistics for a, b and c were focused on obligate and non-obligate complexes separately. Based on previous studies, the area-fraction is computed with two probe sizes, the common radius of 1.4Å (SASA) and the statistically sensible radius of 4.0Å (4ASA).

In summary, four different fraction methods were used for composition data (mole-fraction[full-protein], mole-fraction[surface], area-fraction[1.4Å], and area-fraction[4.0Å]) and 6 different fraction methods for the pairing preferences data (mole-fraction[full-protein], mole-fraction[surface], mole-fraction[interface], contact-fraction, area-fraction[1.4Å], and area-fraction[4.0Å]). In the cases of secondary structure element composition and pairing properties, the area-fraction method was not applied.

6.2.2.3. Amino Acid-Classes

The ability of certain properties of complexes to separate obligate and non-obligate interactions was already analyzed in chapter 4. A great improvement in the sensibility of the predictions was found once amino acid-classes were used instead of the 20 individual amino acid-types. Such classes group several amino acids that share certain properties. Three classification schemes were employed. As shown in table 12, the simplest classification distinguishes the amino acid-types into H-bond forming and non-forming residues (group label 2). The next finer level accounts for the different physicochemical properties and divides the amino acids into hydrophobic, hydrophilic uncharged, negatively and positively charged residues (group label 4). The last level contains small, hydrophobic, negatively and positively charged, and polar amino acids (group label 5).

6.2.2.4. Feature Collection

9,692 features in 347 feature vectors (properties) were collected for a set of 252 obligate and 212 non-obligate protein-protein complexes. Nearly all features are based on the interface region of interacting chains. Most of the data comprise composition and pairing propensity statistics using different interface criteria, fraction methods and amino acid-classes. The features can be divided into 5 sets:

Composition data: Residue, class of residues and secondary structure element compositions at the interface region were counted and converted into statistical potentials using different interface criteria (table 15A). This resulted in 153 feature vectors.

Pairing propensity data: Residue, class of residues and secondary structure element pairing propensities at the interface region were counted and converted into statistical potentials using different interface criteria (table 15B). 160 feature vectors were collected.

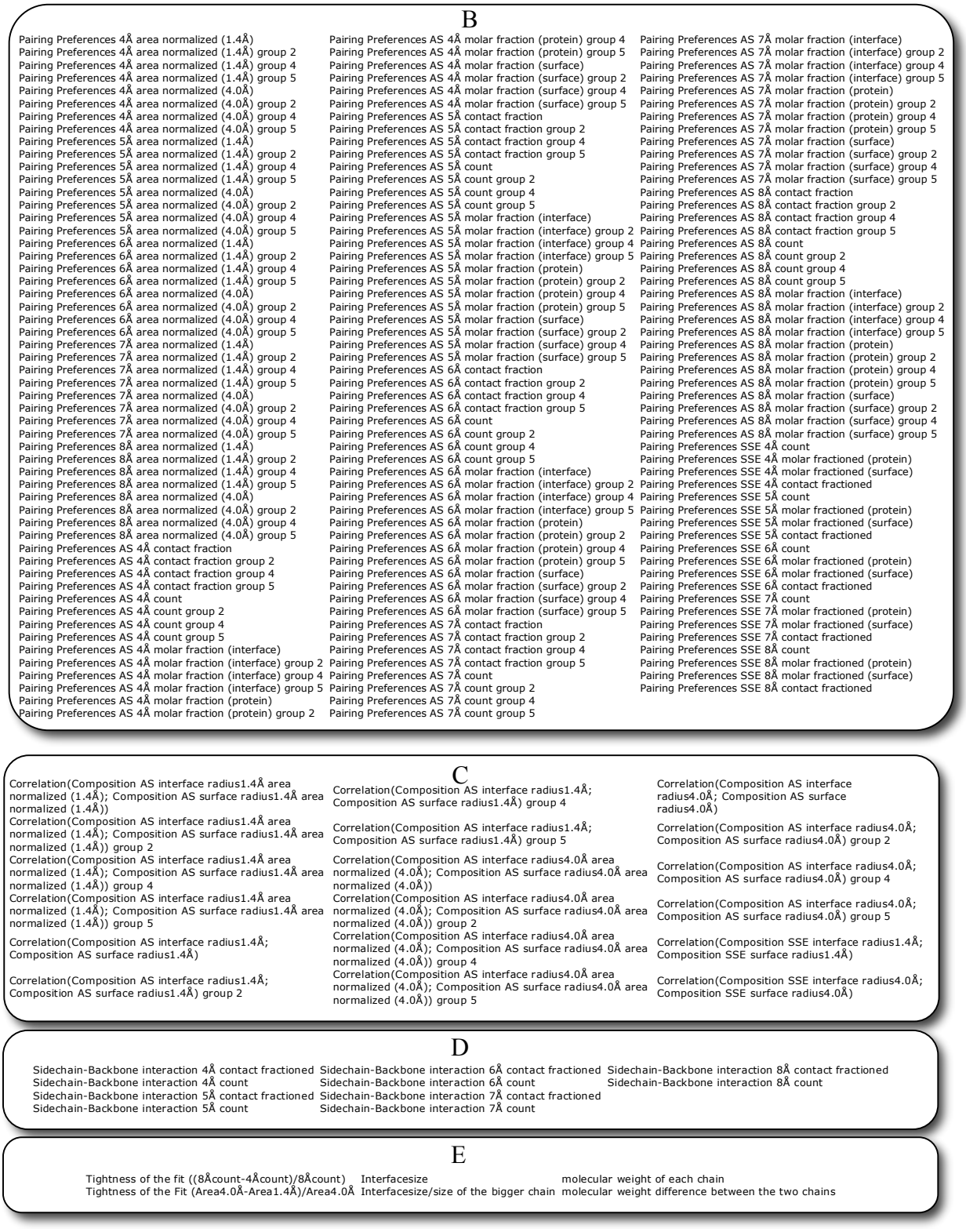
Correlation data: For the residue, residue classes and secondary structure element compositions at the interface and surface, Pearson correlation coefficients were calculated. Changes in accessible surface area for both probe sizes were used to characterize the interface and surface regions where the interface region is a subset of the surface region (table 15C). This led to 18 feature vectors.

Side-chain backbone data: Based on the interaction of heavy atoms among complexed chains, it was analyzed whether an interacting heavy atom belongs to the backbone or to the side-chain of the interacting residue. Using all distance based interface criteria and additionally applying the contact-fraction method, a number of 10 feature vectors were generated (table 15D).

Geometric features: this set includes 6 features (table 15E). The tightness of the fit was defined as the difference between the number of interface residues for the distance criteria 8Å and 4Å divided by the number of interface residues at 8Å distance cutoff. A tight fit would lead to nearly the same number of residues at a distance cutoff of 8Å and 4Å. Dividing by the number of residues gathered at the distance cutoff of 8Å will normalize the results. Another very similar definition for the tightness of the fit was defined as the rate of the interface area size given by a probe with the radius of 1.4Å and the interface area size given by a probe with the radius of 4.0Å. Furthermore, additional features were included that led to a clear distinction between obligate and non-obligate complexes in two related and recently published studies. Zhu et al. reported a successful classification by considering the interface area size and interface area size ratio to the size of the bigger chain in the complex [68]. On the other hand, Kottha et al. found significant differences between obligate and non-obligate complexes when considering the molecular weight of the interacting chains and the difference of molecular weight within the interacting chains [145].

A

Composition AS 4Å area normalized (1.4Å)
Composition AS 4Å area normalized (1.4Å) group 2
Composition AS 4Å area normalized (1.4Å) group 4
Composition AS 4Å area normalized (1.4Å) group 5
Composition AS 4Å area normalized (4.0Å)
Composition AS 4Å area normalized (4.0Å) group 2
Composition AS 4Å area normalized (4.0Å) group 4
Composition AS 4Å area normalized (4.0Å) group 5
Composition AS 4Å count
Composition AS 4Å count group 2
Composition AS 4Å count group 4
Composition AS 4Å count group 5
Composition AS 4Å molar fractioned (protein)
Composition AS 4Å molar fractioned (protein) group 2
Composition AS 4Å molar fractioned (protein) group 4
Composition AS 4Å molar fractioned (protein) group 5
Composition AS 4Å molar fractioned (surface)
Composition AS 4Å molar fractioned (surface) group 2
Composition AS 4Å molar fractioned (surface) group 4
Composition AS 4Å molar fractioned (surface) group 5
Composition AS 5Å area normalized (1.4Å)
Composition AS 5Å area normalized (1.4Å) group 2
Composition AS 5Å area normalized (1.4Å) group 4
Composition AS 5Å area normalized (1.4Å) group 5
Composition AS 5Å area normalized (4.0Å)
Composition AS 5Å area normalized (4.0Å) group 2
Composition AS 5Å area normalized (4.0Å) group 4
Composition AS 5Å area normalized (4.0Å) group 5
Composition AS 5Å count
Composition AS 5Å count group 2
Composition AS 5Å count group 4
Composition AS 5Å count group 5
Composition AS 5Å molar fractioned (protein)
Composition AS 5Å molar fractioned (protein) group 2
Composition AS 5Å molar fractioned (protein) group 4
Composition AS 5Å molar fractioned (protein) group 5
Composition AS 5Å molar fractioned (surface)
Composition AS 5Å molar fractioned (surface) group 2
Composition AS 5Å molar fractioned (surface) group 4
Composition AS 5Å molar fractioned (surface) group 5
Composition AS 6Å area normalized (1.4Å)
Composition AS 6Å area normalized (1.4Å) group 2
Composition AS 6Å area normalized (1.4Å) group 4
Composition AS 6Å area normalized (1.4Å) group 5
Composition AS 6Å area normalized (4.0Å)
Composition AS 6Å area normalized (4.0Å) group 2
Composition AS 6Å area normalized (4.0Å) group 4
Composition AS 6Å area normalized (4.0Å) group 5
Composition AS 6Å count
Composition AS 6Å count group 2
Composition AS 6Å count group 4
Composition AS 6Å count group 5
Composition AS 6Å molar fractioned (protein)
Composition AS 6Å molar fractioned (protein) group 2
Composition AS 6Å molar fractioned (protein) group 4
Composition AS 6Å molar fractioned (protein) group 5
Composition AS 6Å molar fractioned (surface)
Composition AS 6Å molar fractioned (surface) group 2
Composition AS 6Å molar fractioned (surface) group 4
Composition AS 6Å molar fractioned (surface) group 5
Composition AS 7Å area normalized (1.4Å)
Composition AS 7Å area normalized (1.4Å) group 2
Composition AS 7Å area normalized (1.4Å) group 4
Composition AS 7Å area normalized (1.4Å) group 5
Composition AS 7Å area normalized (4.0Å)
Composition AS 7Å area normalized (4.0Å) group 2
Composition AS 7Å area normalized (4.0Å) group 4
Composition AS 7Å area normalized (4.0Å) group 5
Composition AS 7Å count
Composition AS 7Å count group 2
Composition AS 7Å count group 4
Composition AS 7Å count group 5
Composition AS 7Å molar fractioned (protein)
Composition AS 7Å molar fractioned (protein) group 2
Composition AS 7Å molar fractioned (protein) group 4
Composition AS 7Å molar fractioned (protein) group 5
Composition AS 7Å molar fractioned (surface)
Composition AS 7Å molar fractioned (surface) group 2
Composition AS 7Å molar fractioned (surface) group 4
Composition AS 7Å molar fractioned (surface) group 5
Composition AS 8Å area normalized (1.4Å)
Composition AS 8Å area normalized (1.4Å) group 2
Composition AS 8Å area normalized (1.4Å) group 4
Composition AS 8Å area normalized (1.4Å) group 5
Composition AS 8Å area normalized (4.0Å)
Composition AS 8Å area normalized (4.0Å) group 2
Composition AS 8Å area normalized (4.0Å) group 4
Composition AS 8Å area normalized (4.0Å) group 5
Composition AS 8Å count
Composition AS 8Å count group 2
Composition AS 8Å count group 4
Composition AS 8Å count group 5
Composition AS 8Å molar fractioned (protein)
Composition AS 8Å molar fractioned (protein) group 2
Composition AS 8Å molar fractioned (protein) group 4
Composition AS 8Å molar fractioned (protein) group 5
Composition AS 8Å molar fractioned (surface)
Composition AS 8Å molar fractioned (surface) group 2
Composition AS 8Å molar fractioned (surface) group 4
Composition AS 8Å molar fractioned (surface) group 5
Composition AS radius1.4Å area normalized (1.4Å)
Composition AS radius1.4Å area normalized (1.4Å) group 2
Composition AS radius1.4Å area normalized (1.4Å) group 4
Composition AS radius1.4Å area normalized (1.4Å) group 5
Composition AS radius1.4Å count
Composition AS radius1.4Å count group 2
Composition AS radius1.4Å count group 4
Composition AS radius1.4Å count group 5
Composition AS radius1.4Å molar fractioned (protein)
Composition AS radius1.4Å molar fractioned (protein) group 2
Composition AS radius1.4Å molar fractioned (protein) group 4
Composition AS radius1.4Å molar fractioned (protein) group 5
Composition AS radius1.4Å molar fractioned (surface)
Composition AS radius1.4Å molar fractioned (surface) group 2
Composition AS radius1.4Å molar fractioned (surface) group 4
Composition AS radius1.4Å molar fractioned (surface) group 5
Composition AS radius4.0Å area normalized (4.0Å)
Composition AS radius4.0Å area normalized (4.0Å) group 2
Composition AS radius4.0Å area normalized (4.0Å) group 4
Composition AS radius4.0Å area normalized (4.0Å) group 5
Composition AS radius4.0Å count
Composition AS radius4.0Å count group 2
Composition AS radius4.0Å count group 4
Composition AS radius4.0Å count group 5
Composition AS radius4.0Å molar fractioned (protein)
Composition AS radius4.0Å molar fractioned (protein) group 2
Composition AS radius4.0Å molar fractioned (protein) group 4
Composition AS radius4.0Å molar fractioned (protein) group 5
Composition AS radius4.0Å molar fractioned (surface)
Composition AS radius4.0Å molar fractioned (surface) group 2
Composition AS radius4.0Å molar fractioned (surface) group 4
Composition AS radius4.0Å molar fractioned (surface) group 5
Composition SSE 4Å count
Composition SSE 4Å molar fractioned (protein)
Composition SSE 4Å molar fractioned (surface)
Composition SSE 5Å count
Composition SSE 5Å molar fractioned (protein)
Composition SSE 5Å molar fractioned (surface)
Composition SSE 6Å count
Composition SSE 6Å molar fractioned (protein)
Composition SSE 6Å molar fractioned (surface)
Composition SSE 7Å count
Composition SSE 7Å molar fractioned (protein)
Composition SSE 7Å molar fractioned (surface)
Composition SSE 8Å count
Composition SSE 8Å molar fractioned (protein)
Composition SSE 8Å molar fractioned (surface)
Composition SSE radius1.4Å count
Composition SSE radius1.4Å molar fractioned (protein)
Composition SSE radius1.4Å molar fractioned (surface)
Composition SSE radius4.0Å count
Composition SSE radius4.0Å molar fractioned (protein)
Composition SSE radius4.0Å molar fractioned (surface)



B

Pairing Preferences 4Å area normalized (1.4Å)
Pairing Preferences 4Å area normalized (1.4Å) group 2
Pairing Preferences 4Å area normalized (1.4Å) group 4
Pairing Preferences 4Å area normalized (1.4Å) group 5
Pairing Preferences 4Å area normalized (4.0Å)
Pairing Preferences 4Å area normalized (4.0Å) group 2
Pairing Preferences 4Å area normalized (4.0Å) group 4
Pairing Preferences 4Å area normalized (4.0Å) group 5
Pairing Preferences 5Å area normalized (1.4Å)
Pairing Preferences 5Å area normalized (1.4Å) group 2
Pairing Preferences 5Å area normalized (1.4Å) group 4
Pairing Preferences 5Å area normalized (1.4Å) group 5
Pairing Preferences 5Å area normalized (4.0Å)
Pairing Preferences 5Å area normalized (4.0Å) group 2
Pairing Preferences 5Å area normalized (4.0Å) group 4
Pairing Preferences 5Å area normalized (4.0Å) group 5
Pairing Preferences 6Å area normalized (1.4Å)
Pairing Preferences 6Å area normalized (1.4Å) group 2
Pairing Preferences 6Å area normalized (1.4Å) group 4
Pairing Preferences 6Å area normalized (1.4Å) group 5
Pairing Preferences 6Å area normalized (4.0Å)
Pairing Preferences 6Å area normalized (4.0Å) group 2
Pairing Preferences 6Å area normalized (4.0Å) group 4
Pairing Preferences 6Å area normalized (4.0Å) group 5
Pairing Preferences 7Å area normalized (1.4Å)
Pairing Preferences 7Å area normalized (1.4Å) group 2
Pairing Preferences 7Å area normalized (1.4Å) group 4
Pairing Preferences 7Å area normalized (1.4Å) group 5
Pairing Preferences 7Å area normalized (4.0Å)
Pairing Preferences 7Å area normalized (4.0Å) group 2
Pairing Preferences 7Å area normalized (4.0Å) group 4
Pairing Preferences 7Å area normalized (4.0Å) group 5
Pairing Preferences 8Å area normalized (1.4Å)
Pairing Preferences 8Å area normalized (1.4Å) group 2
Pairing Preferences 8Å area normalized (1.4Å) group 4
Pairing Preferences 8Å area normalized (1.4Å) group 5
Pairing Preferences 8Å area normalized (4.0Å)
Pairing Preferences 8Å area normalized (4.0Å) group 2
Pairing Preferences 8Å area normalized (4.0Å) group 4
Pairing Preferences 8Å area normalized (4.0Å) group 5
Pairing Preferences AS 4Å contact fraction
Pairing Preferences AS 4Å contact fraction group 2
Pairing Preferences AS 4Å contact fraction group 4
Pairing Preferences AS 4Å contact fraction group 5
Pairing Preferences AS 4Å count
Pairing Preferences AS 4Å count group 2
Pairing Preferences AS 4Å count group 4
Pairing Preferences AS 4Å count group 5
Pairing Preferences AS 4Å molar fraction (interface)
Pairing Preferences AS 4Å molar fraction (interface) group 2
Pairing Preferences AS 4Å molar fraction (interface) group 4
Pairing Preferences AS 4Å molar fraction (interface) group 5
Pairing Preferences AS 4Å molar fraction (protein)
Pairing Preferences AS 4Å molar fraction (protein) group 2
Pairing Preferences AS 4Å molar fraction (protein) group 4
Pairing Preferences AS 4Å molar fraction (protein) group 5
Pairing Preferences AS 4Å molar fraction (surface)
Pairing Preferences AS 4Å molar fraction (surface) group 2
Pairing Preferences AS 4Å molar fraction (surface) group 4
Pairing Preferences AS 4Å molar fraction (surface) group 5
Pairing Preferences AS 5Å contact fraction
Pairing Preferences AS 5Å contact fraction group 2
Pairing Preferences AS 5Å contact fraction group 4
Pairing Preferences AS 5Å contact fraction group 5
Pairing Preferences AS 5Å count
Pairing Preferences AS 5Å count group 2
Pairing Preferences AS 5Å count group 4
Pairing Preferences AS 5Å count group 5
Pairing Preferences AS 5Å molar fraction (interface)
Pairing Preferences AS 5Å molar fraction (interface) group 2
Pairing Preferences AS 5Å molar fraction (interface) group 4
Pairing Preferences AS 5Å molar fraction (interface) group 5
Pairing Preferences AS 5Å molar fraction (protein)
Pairing Preferences AS 5Å molar fraction (protein) group 2
Pairing Preferences AS 5Å molar fraction (protein) group 4
Pairing Preferences AS 5Å molar fraction (protein) group 5
Pairing Preferences AS 5Å molar fraction (surface)
Pairing Preferences AS 5Å molar fraction (surface) group 2
Pairing Preferences AS 5Å molar fraction (surface) group 4
Pairing Preferences AS 5Å molar fraction (surface) group 5
Pairing Preferences AS 6Å contact fraction
Pairing Preferences AS 6Å contact fraction group 2
Pairing Preferences AS 6Å contact fraction group 4
Pairing Preferences AS 6Å contact fraction group 5
Pairing Preferences AS 6Å count
Pairing Preferences AS 6Å count group 2
Pairing Preferences AS 6Å count group 4
Pairing Preferences AS 6Å count group 5
Pairing Preferences AS 6Å molar fraction (interface)
Pairing Preferences AS 6Å molar fraction (interface) group 2
Pairing Preferences AS 6Å molar fraction (interface) group 4
Pairing Preferences AS 6Å molar fraction (interface) group 5
Pairing Preferences AS 6Å molar fraction (protein)
Pairing Preferences AS 6Å molar fraction (protein) group 2
Pairing Preferences AS 6Å molar fraction (protein) group 4
Pairing Preferences AS 6Å molar fraction (protein) group 5
Pairing Preferences AS 6Å molar fraction (surface)
Pairing Preferences AS 6Å molar fraction (surface) group 2
Pairing Preferences AS 6Å molar fraction (surface) group 4
Pairing Preferences AS 6Å molar fraction (surface) group 5
Pairing Preferences AS 7Å contact fraction
Pairing Preferences AS 7Å contact fraction group 2
Pairing Preferences AS 7Å contact fraction group 4
Pairing Preferences AS 7Å contact fraction group 5
Pairing Preferences AS 7Å count
Pairing Preferences AS 7Å count group 2
Pairing Preferences AS 7Å count group 4
Pairing Preferences AS 7Å count group 5
Pairing Preferences AS 7Å molar fraction (interface)
Pairing Preferences AS 7Å molar fraction (interface) group 2
Pairing Preferences AS 7Å molar fraction (interface) group 4
Pairing Preferences AS 7Å molar fraction (interface) group 5
Pairing Preferences AS 7Å molar fraction (protein)
Pairing Preferences AS 7Å molar fraction (protein) group 2
Pairing Preferences AS 7Å molar fraction (protein) group 4
Pairing Preferences AS 7Å molar fraction (protein) group 5
Pairing Preferences AS 7Å molar fraction (surface)
Pairing Preferences AS 7Å molar fraction (surface) group 2
Pairing Preferences AS 7Å molar fraction (surface) group 4
Pairing Preferences AS 7Å molar fraction (surface) group 5
Pairing Preferences AS 8Å contact fraction
Pairing Preferences AS 8Å contact fraction group 2
Pairing Preferences AS 8Å contact fraction group 4
Pairing Preferences AS 8Å contact fraction group 5
Pairing Preferences AS 8Å count
Pairing Preferences AS 8Å count group 2
Pairing Preferences AS 8Å count group 4
Pairing Preferences AS 8Å count group 5
Pairing Preferences AS 8Å molar fraction (interface)
Pairing Preferences AS 8Å molar fraction (interface) group 2
Pairing Preferences AS 8Å molar fraction (interface) group 4
Pairing Preferences AS 8Å molar fraction (interface) group 5
Pairing Preferences AS 8Å molar fraction (protein)
Pairing Preferences AS 8Å molar fraction (protein) group 2
Pairing Preferences AS 8Å molar fraction (protein) group 4
Pairing Preferences AS 8Å molar fraction (protein) group 5
Pairing Preferences AS 8Å molar fraction (surface)
Pairing Preferences AS 8Å molar fraction (surface) group 2
Pairing Preferences AS 8Å molar fraction (surface) group 4
Pairing Preferences AS 8Å molar fraction (surface) group 5
Pairing Preferences SSE 4Å count
Pairing Preferences SSE 4Å molar fractioned (protein)
Pairing Preferences SSE 4Å molar fractioned (surface)
Pairing Preferences SSE 4Å contact fractioned
Pairing Preferences SSE 5Å count
Pairing Preferences SSE 5Å molar fractioned (protein)
Pairing Preferences SSE 5Å molar fractioned (surface)
Pairing Preferences SSE 5Å contact fractioned
Pairing Preferences SSE 6Å count
Pairing Preferences SSE 6Å molar fractioned (protein)
Pairing Preferences SSE 6Å molar fractioned (surface)
Pairing Preferences SSE 6Å contact fractioned
Pairing Preferences SSE 7Å count
Pairing Preferences SSE 7Å molar fractioned (protein)
Pairing Preferences SSE 7Å molar fractioned (surface)
Pairing Preferences SSE 7Å contact fractioned
Pairing Preferences SSE 8Å count
Pairing Preferences SSE 8Å molar fractioned (protein)
Pairing Preferences SSE 8Å molar fractioned (surface)
Pairing Preferences SSE 8Å contact fractioned

C

Correlation(Composition AS interface radius1.4Å area normalized (1.4Å); Composition AS surface radius1.4Å area normalized (1.4Å))
Correlation(Composition AS interface radius1.4Å area normalized (1.4Å); Composition AS surface radius1.4Å area normalized (1.4Å)) group 2
Correlation(Composition AS interface radius1.4Å area normalized (1.4Å); Composition AS surface radius1.4Å area normalized (1.4Å)) group 4
Correlation(Composition AS interface radius1.4Å area normalized (1.4Å); Composition AS surface radius1.4Å area normalized (1.4Å)) group 5
Correlation(Composition AS interface radius1.4Å; Composition AS surface radius1.4Å)
Correlation(Composition AS interface radius1.4Å; Composition AS surface radius1.4Å) group 2
Correlation(Composition AS interface radius1.4Å; Composition AS surface radius1.4Å) group 4
Correlation(Composition AS interface radius1.4Å; Composition AS surface radius1.4Å) group 5
Correlation(Composition AS interface radius4.0Å area normalized (4.0Å); Composition AS surface radius4.0Å area normalized (4.0Å))
Correlation(Composition AS interface radius4.0Å area normalized (4.0Å); Composition AS surface radius4.0Å area normalized (4.0Å)) group 2
Correlation(Composition AS interface radius4.0Å area normalized (4.0Å); Composition AS surface radius4.0Å area normalized (4.0Å)) group 4
Correlation(Composition AS interface radius4.0Å area normalized (4.0Å); Composition AS surface radius4.0Å area normalized (4.0Å)) group 5
Correlation(Composition AS interface radius4.0Å; Composition AS surface radius4.0Å)
Correlation(Composition AS interface radius4.0Å; Composition AS surface radius4.0Å) group 2
Correlation(Composition AS interface radius4.0Å; Composition AS surface radius4.0Å) group 4
Correlation(Composition AS interface radius4.0Å; Composition AS surface radius4.0Å) group 5
Correlation(Composition SSE interface radius1.4Å; Composition SSE surface radius1.4Å)
Correlation(Composition SSE interface radius4.0Å; Composition SSE surface radius4.0Å)

D

Sidechain-Backbone interaction 4Å contact fractioned
Sidechain-Backbone interaction 4Å count
Sidechain-Backbone interaction 5Å contact fractioned
Sidechain-Backbone interaction 5Å count
Sidechain-Backbone interaction 6Å contact fractioned
Sidechain-Backbone interaction 6Å count
Sidechain-Backbone interaction 7Å contact fractioned
Sidechain-Backbone interaction 7Å count
Sidechain-Backbone interaction 8Å contact fractioned
Sidechain-Backbone interaction 8Å count

E

Tightness of the fit ((8Åcount-4Åcount)/8Åcount)
Tightness of the Fit (Area4.0Å-Area1.4Å)/Area4.0Å
Interfacesize
Interfacesize/size of the bigger chain
molecular weight of each chain
molecular weight difference between the two chains

Table 15: *List of feature vectors. A) Feature vectors based on compositions, B) feature vectors based on pairing preferences, C) feature vectors based on correlation coefficients, D) side-chain – backbone interactions, and E) 6 feature vectors based on the current literature findings*

6.3. Results

9,692 features in 347 feature-vectors of 251 obligate/permanent and 212 non-obligate/transient complexes were investigated with a support vector machine approach. Four different kernels were used to compute a sensitive filter for the separation of obligate and non-obligate complexes. Performing the leave-one-out cross-validation and additionally an average of 10 times 10fold cross-validation each feature vector was trained and validated separately. The accuracy of the prediction was defined as:

$$\text{Accuracy} = \frac{Sum\ \ of\ \ \ correct\ \ \ predictions}{Sum\ \ of\ \ \ total\ \ \ predictions}$$

In total, the calculation for the uncombined feature vectors led to 642,644 training and validation runs (347×463×4). On 20-nodes of a Dual-Xeon processor cluster the calculations took approximately 2 hours of CPU time. The computational time for each feature vector was estimated and the calculations were equally distributed on all 40 CPUs.

6.3.1. Single Feature Vectors

Figure 51A shows that on average the results from the radial basis kernel gave more correct predictions than the other three kernels. This was also observed in the study of Zhu and coworkers [68]. However, looking at the results for the sigmoid kernel function in figure 51A shows a standard deviation of 28.83. With an average number of correct predictions of 263.88 a lower limit of 235.05 can be detected. Considering this limit as a standardized limit, a number of feature vectors must have led to accuracies lower than 50%, which should not result from a support vector machines approach. The worst distribution of data-points can only be equal and lead to an accuracy of 50%. Analyzing the results did not reveal any errors. It was assumed that a number of feature vectors from complexes with very small interface regions may lead to overlapping data-points from different complex types. Such information could be interpreted as biased data that may have reduced the accuracy of 50%. However, as the training and testing set of this approach were the same for all four kernel functions, overlapping data-points must have led to low accuracies for all kernel functions. In fact, such low accurate results were

mostly obtained with the sigmoid and in some cases also with the polynomial kernel function. It is still unclear, why these two kernel functions may result in accuracies lower than 50%. Therefore the results derived from the sigmoid and polynomial kernel functions are no longer discussed. Also, since the radial kernel function scored most accurate and is typically used in the literature, further results are evaluated from the radial kernel function only.

Combining all residue, secondary structure element, and side-chain backbone features (figure 51B), the average accuracy for side-chain backbone data surprisingly showed the highest value in cases where the remaining types of feature vectors were not considered (Rest). The data based on residues clearly has the most statistical strength and therefore may consist of many low scoring (weak accuracies) feature vectors summed in an average number of correct predictions. Therefore, putting too much emphasis on this graph is refrained. The group of 'Rest' is based on the 6 additional low dimensional feature vectors, which are analyzed in the process of this section. A point that was previously addressed but not clearly answered (see chapter 4) was whether the predictability benefits from grouping the amino acid into classes. In principle, reducing the dimension of a property by focusing on biophysical properties of the amino acids appears like a promising approach. However, figure 51C shows that this is not the case. Although the accuracy is increased in going from class 5 to 4 and 2, class 20 using no grouping scores the best. In this analysis, the statistical strength of all four classes is nearly the same. Therefore, one can conclude that unassembled data with large dimensions may lead to clearer separation of obligate and non-obligate complexes. At this point, all results were based on bundled compositional and pairing preferential data together with correlation coefficients. Figure 51D addresses the question of the effectiveness of given data forms on the separation accuracy. Aside the 'Rest' group, compositional and pairing preferential data are nearly equal in accuracy. Here, the small number of vectors based on correlation coefficients that are only one-dimensional feature vectors scored nearly in the range of random separation.

Another interesting point already mentioned few times in previous chapters is, which interface criterion will lead to the best distinction of the complex types and may therefore suit the native interfaces most? Figure 51E compares seven different interface criteria that

are partially common in the literature. The graph does not show clear results. Distance-based interface criteria with a cutoff value of 4Å may lead to weak statistical data as some complexes in the dataset do not contain any interactions within this range threshold. This small number of data may lead to insufficient statistics. This is similar in the case of area-based interface criterion. Previously, it was observed that a probe with the size of 1.4Å radius may confuse buried cavities with surface area. Therefore, most of the interfaces were found to be solvent accessible. When this method is used as an interface criterion, only a small number of tightly buried interaction pairs can be found. This leads to the same problem as for the distance criterion of 4Å and therefore results similar weakly. However, the concept of surface area-loss upon complex formation seems to achieve most accurate results, once buried cavities are not confused. This is the case for a probe with the radius of 4.0Å. As for the remaining distance criteria, no clear trend can be observed.

The last aspect of the feature vectors that is analyzed here is the influence of fraction methods on the predictability of the computed potentials (figure 51F). Clearly, un-fractioned data performed most accurately and leads to another surprising observation.

Considering the previous results allows the assumption that un-grouped residue compositions derived from the area-interface criterion, without applying any fraction method, may lead to the most accurate results. Such kind of a feature vector indeed gave a high accuracy of 74,3% and marks the 7th best accuracy observed within 347 feature vectors. Figure 52 shows the top 10 accuracies within all 347 feature vectors for the radial kernel function. Ranks 2 and 4 to 10 consist of un-grouped residue pairing propensities at different distance cutoffs for interface residues and fraction methods. Aside the results for the fraction methods, these ranks are in good agreement with figures 51B, C, D, and E when also considering the side-chain backbone pairing propensities achieving the second rank. The feature vector “weightabsAB” takes the 3rd rank. This feature vector includes the molecular weights in Dalton of the two interacting chains in the complex sorted after their size. Figure 53 shows two qualities of obligate and non-obligate complexes when considering their molecular weight. First, obligate complexes tend to consist of bigger chains than non-obligate complexes. Second, the molecular weight differences of the two interacting chains are smaller for obligate complexes as

most data-points are collected on the diagonal although the obligate complexes in the dataset only consist of 2.4% of homodimers. This strongly agrees with the findings of Kottha and Schröder [145]. The authors performed a support vector machines approach on 161 permanent and 242 transient complexes. Using the molecular weight difference alone as a feature vector achieved an accuracy of up to 80%. On the current dataset, this property gave an accuracy of 70.19%. Considering that the databases are most likely not the same yet similar, the agreement is acceptable.

The best accuracy was obtained with the feature vector "SBPrefcf37" or "Sidechain-Backbone interaction 6Å contact fractioned" in table 15E and resulted in 74.95% accuracy. Figure 54 reveals a significant difference in the backbone – backbone interaction scores which are averaged over all obligate and non-obligate complexes.

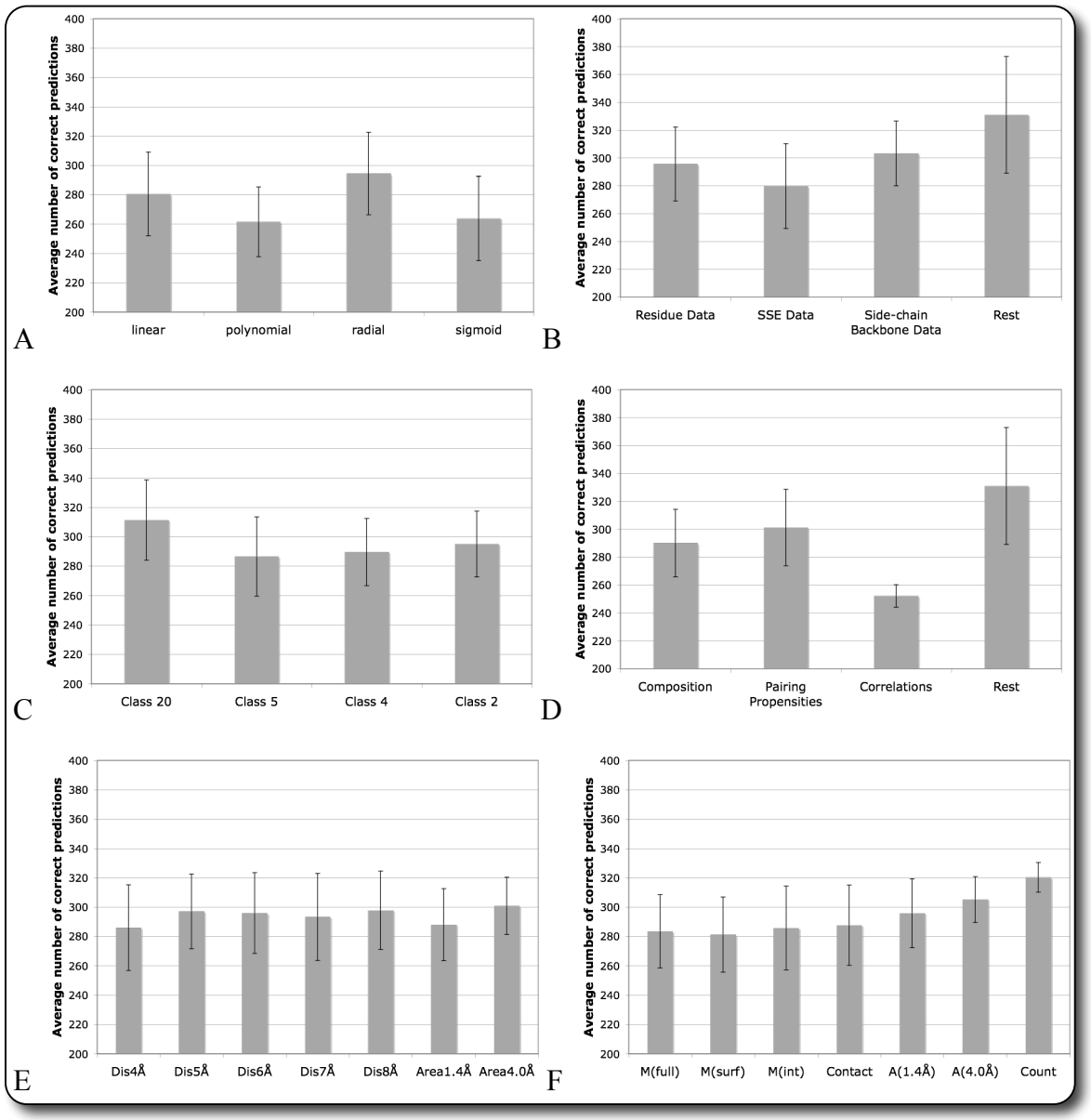


Figure 51: *Average number of correct predictions. 463 is 100% accuracy and 231.5 is 50% accuracy and therefore purely random. (A) The average results based on the kernel function are shown. (B) The average correct predictions based on different types of data for the kernel function radial are shown. 'SSE' = secondary structure elements and 'Rest' contains feature vectors such as weight difference between chains and weight of the chains, tightness related to number of amino acids at the interface and related to the interface area size, the relative size of the interface and the absolute size. (C) The performance of the amino acid-classes based on the radial kernel function is shown. (D) The results of different types of feature vectors are compared for the kernel function radial. 'Rest' contains the same feature vectors as described in B. (E) Compares the performance of different interface criteria for the kernel function radial. 'Dis' refers to distance-based criteria and their cutoff value and 'Area' refers to accessible solvent area criteria based on probe sizes with different radii. (F) Shows the results of the different fraction methods. 'M' stands for mole-fraction methods based on the full protein*

statistics, the surface area statistics, and the interface area statistics. 'Contact' refers to the contact-fraction method and 'A' leads to the area normalization for given probe sizes with the radii 1.4Å and 4.0Å. Finally potentials without fraction methods have also been evaluated in 'Count'. The error bar is based on the standard deviation.

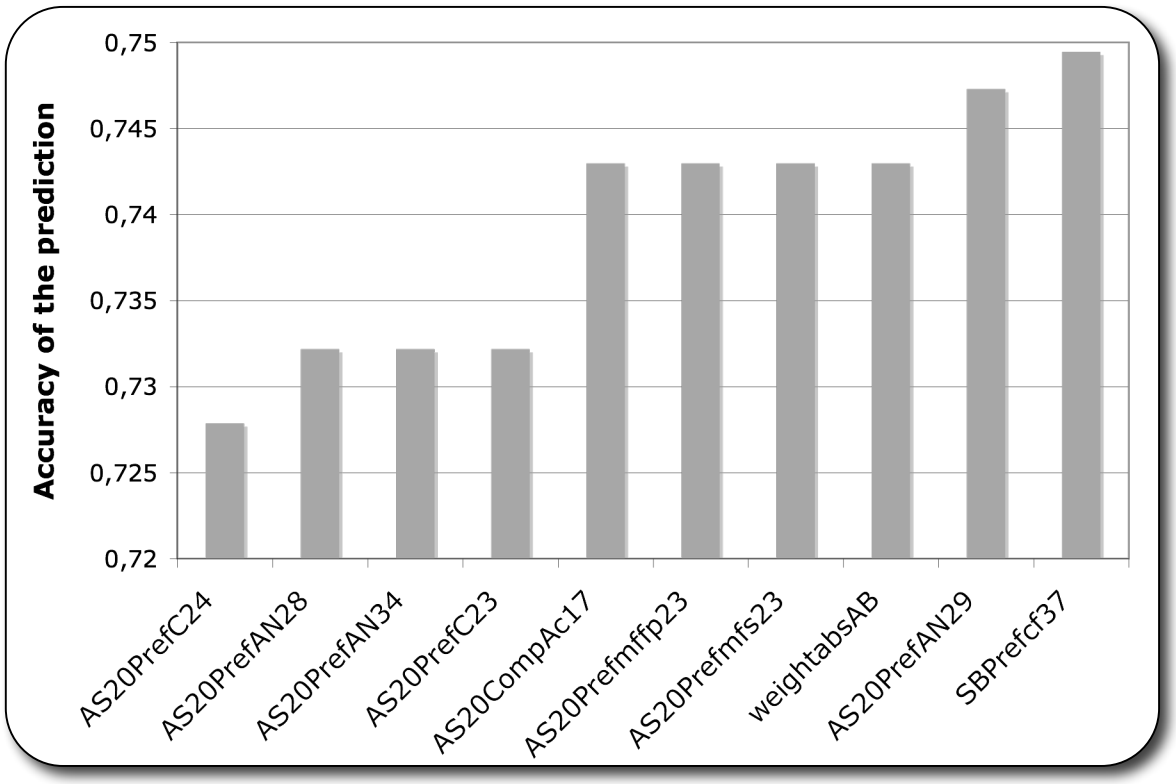


Figure 52*: Top 10 results for the radial kernel function based on their accuracies.* *AS20PrefC24=residue pairing propensities at a distance cutoff of 8Å un-normalized.*
AS20PrefAN28=residue pairing propensities at a distance cutoff of 7Å area-normalized after a probe with the radius of 1.4Å.
AS20PrefAN34=residue pairing propensities at a distance cutoff of 8Å area-normalized after a probe with the radius of 4.0Å.
AS20PrefC23=residue pairing propensities at a distance cutoff of 7Å un-normalized
AS20CompAc17=residue composition of changing surface area upon complex formation based on a probe with the radius of 4.0Å and un-normalized.
AS20Prefmffp23= residue pairing propensities at a distance cutoff of 7Å mole-fractioned after full protein composition.
AS20Prefmfs23=residue pairing propensities at a distance cutoff of 7Å mole-fractioned after surface composition.
weightabsAB=weights of the two interacting chains.
AS20PrefAN29=residue pairing propensities at a distance cutoff of 8Å area-normalized after a probe with the radius of 1.4Å.
SBPrefcf37=side-chain backbone pairing propensities at a distance cutoff of 6Å contact-fractioned.

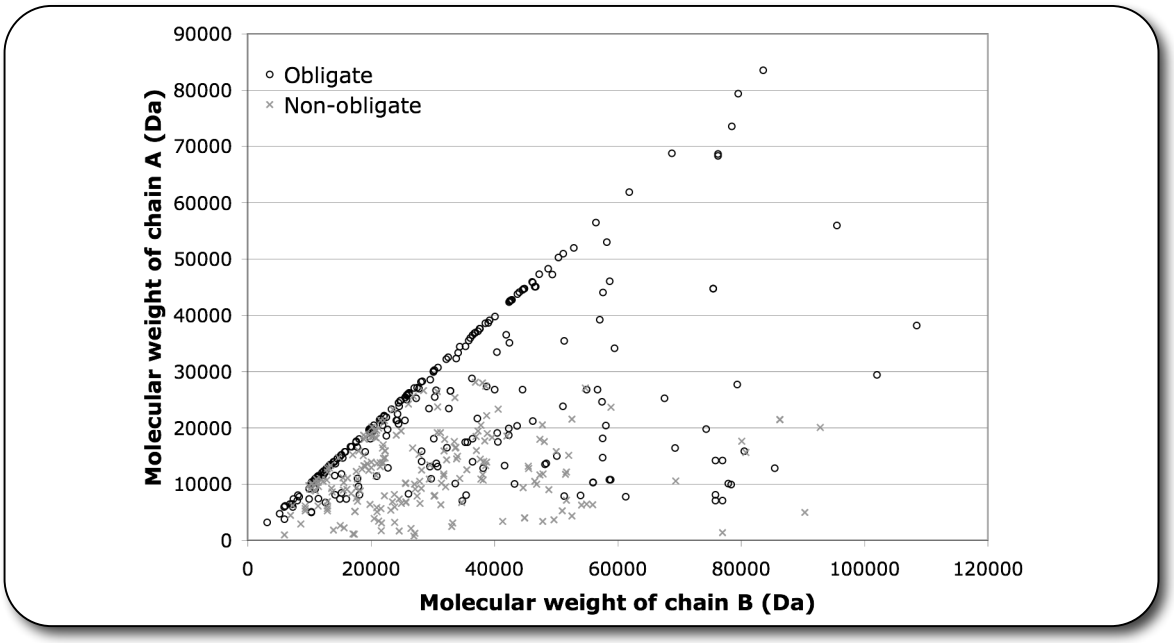


Figure 53: *Plot of the feature vector "weightabsAB" in figure 52 or "molecular weight of each chain" in table 15E.*

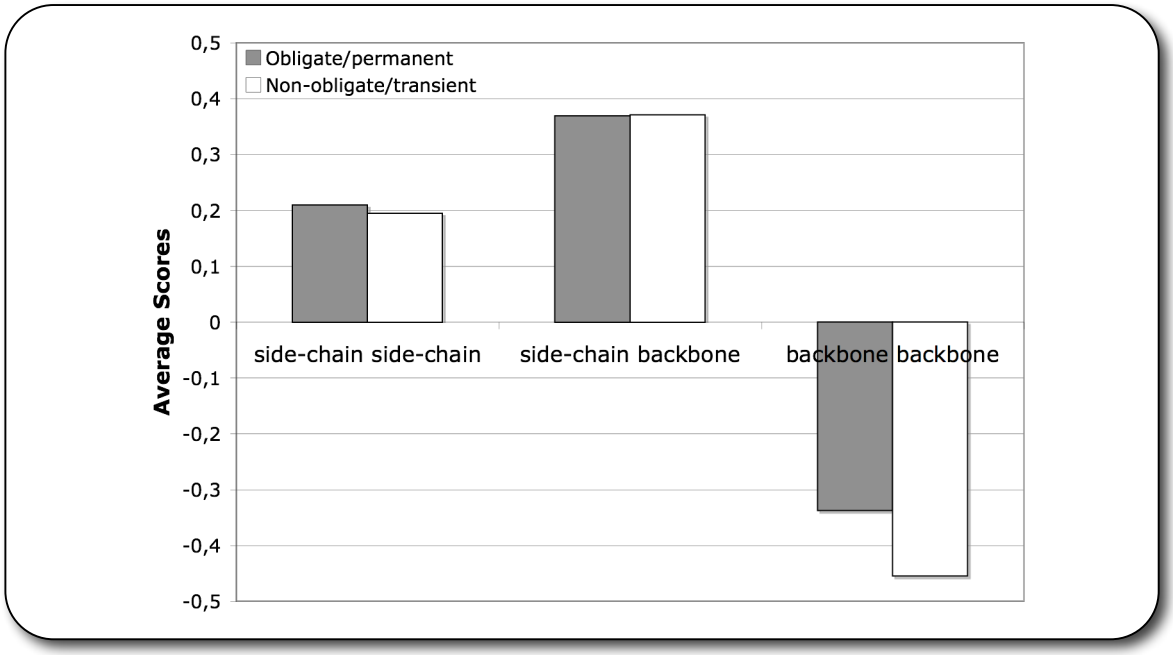


Figure 54: *Plot of the feature vector "SBPrefcf37" in figure 52 or "Sidechain-Backbone interaction 6Å contact fractioned" in table 15E. Average scores for each type of interaction are shown for both types of interfaces.*

6.3.2. Combined Feature Vectors

In the literature accuracies of up to 93.6% were achieved using machine learning approaches and combinations of features [137]. In the previous section, the feature vector "Sidechain-Backbone interaction 6Å contact fractioned" led to an accuracy of 74.95%. Combining this feature vector with all remaining 346 vectors, the risk of falling into a local minimum exists while the computational complexity in finding a good yet not the best combination of feature vectors is low. Combining two feature vectors and using the radial kernel function, an increased accuracy of the "Sidechain-Backbone interaction 6Å contact fractioned" by 0.86% to 75.81% was observed. This accuracy was obtained when combining "Sidechain-Backbone interaction 6Å contact fractioned" with "molecular weight of each chain".

The combination with a third feature vector resulted in an accuracy of 80.78% (374 correct predictions out of 463 predictions). Interestingly, the third feature vector differs from the "Sidechain-Backbone interaction 6Å contact fractioned" only in the interface criterion, which is 8Å. Table 16 shows the results from the leave-one-out cross-validation.

		Predicted		
		Obligate	Non-obligate	**Total**
True	Obligate	208	43	251
	Non-oblgiate	46	166	212
	Total	254	209	463

Table 16: *Leave-one-out cross-validation results for the 3 feature vector-combination of "Sidechain-Backbone interaction 6Å contact fractioned" – "molecular weight of each chain" – "Sidechain-Backbone interaction 8Å contact fractioned" based on 463 predictions.*

As table 16 shows, the prediction accuracies are 82.87% for obligate and 78.30% for non-obligate complexes. In addition to the number of correct predictions, the decision values

for the SVM judging on predictions were evaluated as well. Table 17 shows the results on the average decision values. The region from -0.4524 to 0.4716 indicates the overlapping area for the separation.

	Average decision values
True obligate	0.9457
False obligate	-0.4524
True non-obligate	-0.8147
False non-obligate	0.4716

Table 17: *Average decision values for the leave-one-out cross-validation results based on the 3 feature vector-combination of "Sidechain-Backbone interaction 6Å contact fractioned" – "molecular weight of each chain" – "Sidechain-Backbone interaction 8Å contact fractioned".*

Combining up to 7 feature vectors did not improve the mentioned accuracy of 80.78%, which is in agreement with the literature [145][68] where the highest accessible accuracies were achieved by the combination of only a few features.
Additionally, the relative area size of the interface with area-fractioned amino acid-composition, as suggested by Zhu et al., was also tested. While Zhu and coworkers obtained an accuracy of 88.32%, only 66.74% were achieved in this work.

6.4. Discussion

9,692 features of protein-protein complexes were collected in order to distinguish obligate from non-obligate complexes in a dataset of 463 structures. Mainly focusing on the properties of the interface area, all features were grouped into 347 feature vectors mainly based on composition and pairing propensities of residues and secondary structure elements retrieved from different interface criteria and by applying a number of fraction methods. The R package e1071 interfacing to libsvm was used to perform the support vector machine classification and gave an accuracy of 80.78%. This accuracy was

achieved with combining two very similar pairing propensities of side-chain and backbone combinations and the weights of the two interacting chains. Obviously, mainly the backbone – backbone interactions of the two similar feature vectors lead to the good separation of the obligate and non-obligate dataset. This may support the idea of tight packing in obligate complexes, as the average score for backbone – backbone interactions is significantly higher than in non-obligate complexes. As stated before, the absolute weights may be a great separating aspect as well. Also it was concluded, that the rather low dimensionality of features might lead to better separations of obligate and non-obligate complexes. The current work supports this assumption as the combination of three low dimensional feature vectors resulted in the highest accuracy.

Previously, it was also found that the best separation of obligate and non-obligate complexes could be achieved when considering the pairing propensities of secondary structure elements. In the uncombined trainings and evaluations, the secondary structure element-pairing propensities led to an accuracy of up to 71.92%, which ranks within the top 25 of 347 feature vectors. The results of the current and previous work are therefore in good agreement.

Two features, previously emphasized in the literature, led to lower yet acceptable separations. Kottha et al. found high separation sensitivity for the molecular weight difference of the two interacting chains alone. Using a radial base kernel function with support vector machines, this feature achieved an accuracy of 80%. Applying the same conditions to this dataset led to an accuracy of 70.19%. Although this accuracy is lower it remains surprising that this rather simple feature alone achieves such sensitivity.

As mentioned above, the entire combinatorial space was not considered and a large risk for a local minimum is present. However, an accuracy of 80.78% in distinguishing obligate from non-obligate complexes obtained in a large dataset of 463 structures seems feasible enough for finding more database entries from the large pool of structures in the RCSB PDB and for characterizing interactions for re-evaluation of docking samples, as addressed in chapter 3.

7 Outlook

This thesis introduced and applied several computational methods for analyzing protein-protein interfaces. Statistical analyses, as applied in this work, strongly benefit from a large and clean dataset. The rich informational content of PDB structures has become an essential part of most analyses in the current literature. Although the RCSB PDB contains more than 40.000 structures, only 534 were collected in chapter 4 in order to retrieve additional information. An automated and easy to apply procedure to retrieve suitable protein structures will therefore be one of the most important steps in the further process of chapter 4. With the knowledge gained in chapter 6 such an automated procedure may be easy to develop. Collecting all multichain complexes from the RCSB PDB, three types of interactions may occur: 1. The packing of two chains is of non-obligate/transient or of 2. obligate/permanent nature or 3. it is only a crystal packing. As the separation in non-obligate/transient and obligate/permanent already achieved a satisfying accuracy, another filter should be developed to separate crystal packing from natural complexes. With these filters an automated update function in the ABC database developed in chapter 5 may be an easy implementation, as the program language R will soon be implemented in the database. This will utilize the use of powerful statistical learning approaches such as support vector machines.

An increased dataset in the database will urge the need to define and find data redundancies. The implementation of BioJAVA in the database will allow sequence alignments and the assessment of sequence similarity and homology. This may be used to define redundancies. Additionally, an alternative definition, as used in chapter 6, may be implemented as well, where correlation coefficients for some interface or protein properties may be computed and used to define similarities.

With an increased number of data in the database newer scoring matrices can be calculated. Combining this with the information from chapter 6 where a large number of interface properties were analyzed and a combination of interface and protein properties

was found to lead to clear distinctions of obligate/permanent and non-obligate/transient interactions. A more enhanced scoring function may be applied to rigid-body docking to increase the predictability of the native structure. Instead of scoring residue and secondary structure element-propensities at given distance cutoffs gathered from all proposed docking orientations, one would now focus on the interactions of side-chain atoms and such in the backbone of the amino acids, as defined in chapter 6.

The most focus will therefore be put on the new ABC database. Implementing even the VMD script that is used to gather the rich data from the protein complex structures into the database may lead to a fully automated database updating the data upon each new RCSB PDB entry. A superficial estimation led to nearly 10,000 potential complexes of either obligate/permanent or non-obligate/transient interactions. Such large data may lead to clearer patterns and deeper understanding of protein-protein interactions. A powerful docking and scoring approach may also generate a large number of new complex structures.

Furthermore, different types of interactions may be found. By mainly applying the methods of chapter 4, dendrograms may visualize protein-protein interaction types beyond obligate or non-obligate, and transient or permanent.

Literature

[1] Alm E, Arkin AP. *Biological networks*.
Curr Opin Struct Biol. **2003** Apr;13(2):193-202.

[2] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions*.
Nucleic Acids Res. **2002** Jan 1;30(1):303-5.

[3] Bonafe CF, Matsukuma AY, Matsuura MS. *ATP-induced tetramerization and cooperativity in hemoglobin of lower vertebrates*.
J Biol Chem. **1999** Jan 15;274(3):1196-8.

[4] Chan NL, Rogers PH, Arnone A. *Crystal structure of the S-nitroso form of liganded human hemoglobin.*
Biochemistry. **1998** Nov 24;37(47):16459-64.

[5] Renatus M, Stennicke HR, Scott FL, Liddington RC, Salvesen GS. *Dimer formation drives the activation of the cell death protease caspase 9.*
Proc Natl Acad Sci U S A. **2001** Dec 4;98(25):14250-5.

[6] Jiang G, den Hertog J, Hunter T. *Receptor-like protein tyrosine phosphatase alpha homodimerizes on the cell surface.*
Mol Cell Biol. **2000** Aug;20(16):5917-29.

[7] Bilwes AM, den Hertog J, Hunter T, Noel JP. *Structural basis for inhibition of receptor protein-tyrosine phosphatase-alpha by dimerization.*
Nature. **1996** Aug 8;382(6591):555-9.

[8] Hebert TE, Bouvier M. *Structural and functional aspects of G protein-coupled receptor oligomerization.*
Biochem Cell Biol. **1998**;76(1):1-11.

[9] Rios CD, Jordan BA, Gomes I, Devi LA. *G-protein-coupled receptor dimerization: modulation of receptor function.*
Pharmacol Ther. **2001** Nov-Dec;92(2-3):71-87.

[10] Fotiadis D, Liang Y, Filipek S, Saperstein DA, Engel A, Palczewski K. *Atomic-force microscopy: Rhodopsin dimers in native disc membranes.* *Nature.* **2003** Jan 9;421(6919):127-8.

[11] Kunishima N, Shimada Y, Tsuji Y, Sato T, Yamamoto M, Kumasaka T, Nakanishi S,Jingami H, Morikawa K. *Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor.* *Nature.* **2000** Oct 26;407(6807):971-7.

[12] Breitwieser GE. *G protein-coupled receptor oligomerization: implications for G protein activation and cell signaling.* *Circ Res.* **2004** Jan 9;94(1):17-27.

[13] Job D, Valiron O, Oakley B. *Microtubule nucleation.* *Curr Opin Cell Biol.* **2003** Feb;15(1):111-7.

[14] Emsley J, Knight CG, Farndale RW, Barnes MJ, Liddington RC. *Structural basis of collagen recognition by integrin alpha2beta1.* *Cell.* **2000** Mar 31;101(1):47-56.

[15] Nogales E, Wolf SG, Downing KH. *Structure of the alpha beta tubulin dimer by electron crystallography.* *Nature.* **1998** Jan 8;391(6663):199-203.

[16] Nooren IM, Thornton JM. *Diversity of protein-protein interactions.* *EMBO J.* **2003** Jul 15;22(14):3486-92.

[17] Ofran Y, Rost B. *Analysing six types of protein-protein interfaces.* *J Mol Biol.* **2003** Jan 10;325(2):377-87.

[18] Tsai CJ, Xu D, Nussinov R. *Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes.* *Protein Sci* **1997**; 6:1793-1805.

[19] Chothia C, Janin J. *Principles of protein-protein recognition.* *Nature.* **1975** Aug 28;256(5520):705-8.

[20] Miller S, Lesk AM, Janin J, Chothia C. *The accessible surface area and stability of oligomeric proteins.* *Nature.* **1987** Aug 27-Sep 2;328(6133):834-6.

[21] Argos P. *An investigation of protein subunit and domain interfaces.* *Protein Eng.* **1988** Jul;2(2):101-13.

[22] Janin J, Miller S, Chothia C. *Surface, subunit interfaces and interior of oligomeric proteins.*
J Mol Biol. **1988** Nov 5;204(1):155-64.

[23] Jones S, Thornton JM. *Protein-protein interactions: a review of protein dimer structures.*
Prog Biophys Mol Biol. **1995**;63(1):31-65.

[24] Jones S, Thornton JM. *Principles of protein-protein interactions.*
Proc Natl Acad Sci U S A. **1996** Jan 9;93(1):13-20.

[25] Lo Conte L, Chothia C, Janin J. *The atomic structure of protein-protein recognition sites.*
J Mol Biol. **1999** Feb 5;285(5):2177-98.

[26] Available at **http://www.ebi.ac.uk/embl/index.html**

[27] Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, Baldwin A, Bates K,Bhattacharyya S, Browne P, van den Broek A, Castro M, Duggan K, Eberhardt R,Faruque N, Gamble J, Kanz C, Kulikova T, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, McHale M, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Sobhany S, Stoehr P, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R. *EMBL Nucleotide Sequence Database: developments in 2005.*
Nucleic Acids Res. **2006** Jan 1;34(Database issue):D10-5.

[28] Available at **http://www.ddbj.nig.ac.jp**

[29] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. *GenBank.*
Nucleic Acids Res. **2006** Jan 1;34(Database issue):D16-20.

[30] Sanger F, Nicklen S, Coulson AR. *DNA sequencing with chain-terminating inhibitors.*
Proc Natl Acad Sci U S A. **1977** Dec;74(12):5463-7.

[31] Source: **http://www.ebi.ac.uk/embl**

[32] Watanabe K, Harayama S. *SWISS-PROT: The curated protein sequence database on Internet (in Japanese).*
Protein, Nucleic Acid and Enzyme **2001** 46:80-86.

[33] Source: **http://www.expasy.org/sprot/relnotes/relstat.html**

[34] Tatusov RL, Koonin EV, Lipman DJ. *A genomic perspective on protein families.*
Science. **1997** Oct 24;278(5338):631-7.

[35] Patny A, Desai PV, Avery MA. *Ligand-supported homology modeling of the human angiotensin II type 1 (AT(1)) receptor: Insights into the molecular determinants of telmisartan binding.*
Proteins. **2006** Oct 10; [Epub ahead of print].

[36] Purta E, van Vliet F, Tricot C, De Bie LG, Feder M, Skowronek K, Droogmans L, Bujnicki JM. *Sequence-structure-function relationships of a tRNA (m7G46) methyltransferase studied by homology modeling and site-directed mutagenesis.*
Proteins. **2005** May 15;59(3):482-8.

[37] Allorge D, Breant D, Harlow J, Chowdry J, Lo-Guidice JM, Chevalier D, Cauffiez C, Lhermitte M, Blaney FE, Tucker GT, Broly F, Ellis SW. *Functional analysis of CYP2D6.31 variant: homology modeling suggests possible disruption of redox partner interaction by Arg440His substitution.*
Proteins. **2005** May 1;59(2):339-46.

[38] Reeves DC, Sayed MF, Chau PL, Price KL, Lummis SC. *Prediction of 5-HT3 receptor agonist-binding residues using homology modeling.*
Biophys J. **2003** Apr;84(4):2338-44.

[39] Swalla BM, Gumport RI, Gardner JF. *Conservation of structure and function among tyrosine recombinases: homology-based modeling of the lambda integrase core-binding domain.*
Nucleic Acids Res. **2003** Feb 1;31(3):805-18.

[40] Ogawa H, Toyoshima C. *Homology modeling of the cation binding sites of Na+K+-ATPase.*
Proc Natl Acad Sci U S A. **2002** Dec 10;99(25):15977-82. Epub 2002 Dec 2.

[41] Payne VA, Chang YT, Loew GH. *Homology modeling and substrate binding study of human CYP2C18 and CYP2C19 enzymes.*
Proteins. **1999** Nov 1;37(2):204-17.

[42] Source: **http://www.pdb.org/pdb/statistics**

[43] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. *The Protein Data Bank: a computer-based archival file for macromolecular structures.*
J Mol Biol. **1977** May 25;112(3):535-42.

[44] Available at **http://www.ebi.ac.uk/uniparc/**

[45] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. *CATH--a hierarchic classification of protein domain structures.*
Structure. **1997** Aug 15;5(8):1093-108.

[46] Murzin AG, Brenner SE, Hubbard T, Chothia C. *SCOP: a structural classification of proteins database for the investigation of sequences and structures.*
J Mol Biol. **1995** Apr 7;247(4):536-40.

[47] Doolittle RF. *Similar amino acid sequences: chance or common ancestry?*
Science. **1981** Oct 9;214(4517):149-59.

[48] Chothia C, Lesk AM. *The relation between the divergence of sequence and structure in proteins.*
EMBO J. **1986** Apr;5(4):823-6.

[49] Henikoff S, Henikoff JG. *Performance evaluation of amino acid substitution matrices.*
Proteins. **1993** Sep;17(1):49-61.

[50] Wilbur WJ. *On the PAM matrix model of protein evolution.*
Mol Biol Evol. **1985** Sep;2(5):434-47.

[51] Valencia A, Kjeldgaard M, Pai EF, Sander C. *GTPase domains of ras p21 oncogene protein and elongation factor Tu: analysis of three-dimensional structures, sequence families, and functional sites.*
Proc Natl Acad Sci U S A. **1991** Jun 15;88(12):5443-7.

[52] Holmes KC, Sander C, Valencia A. *A new ATP-binding fold in actin, hexokinase and Hsc70.*
Trends Cell Biol. **1993** Feb;3(2):53-9.

[53] Brenner SE, Chothia C, Hubbard TJ, Murzin AG. *Understanding protein structure: using scop for fold interpretation.*
Methods Enzymol. **1996**;266:635-43.

[54] Sander C, Schneider R. *Database of homology-derived protein structures and the structural meaning of sequence alignment.*
Proteins. **1991**;9(1):56-68.

[55] Needleman SB, Wunsch CD. *A general method applicable to the search for similarities in the amino acid sequence of two proteins.*
J Mol Biol. **1970** Mar;48(3):443-53.

[56] Smith TF, Waterman MS. *Identification of common molecular subsequences.*
J Mol Biol. **1981** Mar 25;147(1):195-7.

[57] Pearson WR, Lipman DJ. *Improved tools for biological sequence comparison.*
Proc Natl Acad Sci U S A. **1988** Apr;85(8):2444-8.

[58] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. *Basic local alignment search tool.*
J Mol Biol. **1990** Oct 5;215(3):403-10.

[59] Karlin S, Altschul SF. *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.*
Proc Natl Acad Sci U S A. **1990** Mar;87(6):2264-8.

[60] Higgins DG, Sharp PM. *CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.*
Gene. **1988** Dec 15;73(1):237-44.

[61] Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. *ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information.*
Bioinformatics. **2003** Jan;19(1):163-4.

[62] Koike A, Takagi T. *Prediction of protein-protein interaction sites using support vector machines.*
Protein Eng Des Sel. **2004** Feb;17(2):165-73. Epub 2004 Jan 20.

[63] Res I, Mihalek I, Lichtarge O. *An evolution based classifier for prediction of protein interfaces without using protein structures.*
Bioinformatics. **2005** May 15;21(10):2496-501. Epub 2005 Feb 22.

[64] Chung JL, Wang W, Bourne PE. *Exploiting sequence and structure homologs to identify protein-protein binding sites.*
Proteins. **2006** Mar 15;62(3):630-40.

[65] Humphrey W, Dalke A, Schulten K. *VMD: visual molecular dynamics.*
J Mol Graph. **1996** Feb;14(1):33-8, 27-8.

[66] Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. *Residue frequencies and pairing preferences at protein-protein interfaces.*
Proteins **2001**; 43:89-102.

[67] Ansari S, Helms V. *Statistical Analysis of Predominantly Transient Protein-Protein Interfaces.*
Proteins **2005**; 61:344-355.

[68] Zhu H, S Domingues F, Sommer I, Lengauer T. *NOXClass: prediction of protein-protein interaction types.*
BMC Bioinformatics. **2006** Jan 19;7:27.

[69] Mintseris J, Weng Z. *Atomic contact vectors in protein-protein recognition.*
Proteins. **2003** Nov 15;53(3):629-39.

[70] Aloy P, Russell RB. *Interrogating protein interaction networks through structural biology.*
Proc Natl Acad Sci U S A. **2002** Apr 30;99(9):5896-901. Epub 2002 Apr 23.

[71] Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. *Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions.*
Protein Sci. **1998**;7:2578-2586.

[72] Bahadur RP, Chakrabarti P, Rodier F, Janin J. *A dissection of specific and non-specific protein-protein interfaces.*
J Mol Biol. **2004** Feb 27;336(4):943-55.

[73] Hubbard SJ, Thornton JM. *NACCESS, Computer Program.*
Department of Biochemistry and Molecular Biology, University College London **1993**.

[74] Lee B, Richards FM. *The interpretation of protein structures: Estimation of static accessibility.*
J Mol Biol. **1971** Feb 14; 55(3): 379-380.

[75] 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 U S A. **1992** Mar 15;89(6):2195-9.

[76] Kohlbacher O, Lenhof HP. *BALL--rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library.*
Bioinformatics. **2000** Sep;16(9):815-24.

[77] Duan Y, Reddy BV, Kaznessis YN. *Physicochemical and residue conservation calculations to improve the ranking of protein-protein docking solutions.*
Protein Sci. **2005** Feb;14(2):316-28.

[78] Murphy J, Gatchell DW, Prasad JC, Vajda S. *Combination of scoring functions improves discrimination in protein-protein docking.*
Proteins. **2003** Dec 1;53(4):840-54.

[79] Moont G, Gabb HA, Sternberg MJ. *Use of pair potentials across protein interfaces in screening predicted docked complexes.*
Proteins. **1999** May 15;35(3):364-73.

[80] Source: **http://www.expasy.org/tools/pscale/A.A.Swiss-Prot.html**

[81] Jain AK, Murty MN, Flynn PJ. *Data clustering: a review.*
ACM Computing Surveys. **1999** Sep 1;31(3):264-323.

[82] Saitou N, Nei M. *The neighbor-joining method: a new method for reconstructing phylogenetic trees.*
Mol Biol Evol. **1987** Jul;4(4):406-25.

[83] Siddiqui AS, Dengler U, Barton GJ. *3Dee: a database of protein structural domains.*
Bioinformatics. **2001** Feb;17(2):200-1.

[84] Sowdhamini R, Burke DF, Huang JF, Mizuguchi K, Nagarajaram HA, Srinivasan N, Steward RE, Blundell TL. *CAMPASS: a database of structurally aligned protein superfamilies.*
Structure. **1998** Sep 15;6(9):1087-94.

[85] Shindyalov IN, Bourne PE. *A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm.*
Nucleic Acids Res. **2001** Jan 1;29(1):228-9.

[86] Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. *A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3.*
Nucleic Acids Res. **2001** Jan 1;29(1):55-7.

[87] Bray JE, Todd AE, Pearl FM, Thornton JM, Orengo CA. *The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues.*
Protein Eng. **2000** Mar;13(3):153-65.

[88] Wang Y, Addess KJ, Geer L, Madej T, Marchler-Bauer A, Zimmerman D, Bryant SH. *MMDB: 3D structure data in Entrez.*
Nucleic Acids Res. **2000** Jan 1;28(1):243-5.

[89] Holm L, Sander C. *The FSSP database: fold classification based on structure-structure alignment of proteins.*
Nucleic Acids Res. **1996** Jan 1;24(1):206-9.

[90] Mizuguchi K, Deane CM, Blundell TL, Overington JP. *HOMSTRAD: a database of protein structure alignments for homologous families.*
Protein Sci. **1998** Nov;7(11):2469-71.

[91] Jones S, Thornton JM. *Analysis of protein-protein interaction sites using surface patches.*
J Mol Biol. **1997** Sep 12;272(1):121-32.

[92] McCoy AJ, Chandana Epa V, Colman PM. *Electrostatic complementarity at protein/protein interfaces.*
J Mol Biol. **1997** May 2;268(2):570-84.

[93] Sheinerman FB, Norel R, Honig B. *Electrostatic aspects of protein-protein interactions.*
Curr Opin Struct Biol. **2000** Apr;10(2):153-9.

[94] Jones S, Marin A, Thornton JM. *Protein domain interfaces: characterization and comparison with oligomeric protein interfaces.*
Protein Eng. **2000** Feb;13(2):77-82.

[95] Sternberg MJ, Gabb HA, Jackson RM. *Predictive docking of protein-protein and protein-DNA complexes.*
Curr Opin Struct Biol. **1998** Apr;8(2):250-6.

[96] Ponder JW, Richards FM. *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.*
J Mol Biol. **1987** Feb 20;193(4):775-91.

[97] Hubbard SJ, Argos P. *Cavities and packing at protein interfaces.*
Protein Sci. **1994** Dec;3(12):2194-206.

[98] Jiang S, Tovchigrechko A, Vakser IA. *The role of geometric complementarity in secondary structure packing: a systematic docking study.*
Protein Sci. **2003** Aug;12(8):1646-51.

[99] Richards FM. *Areas, volumes, packing and protein structure.*
Annu Rev Biophys Bioeng. **1977**;6:151-76.

[100] Lawrence MC, Colman PM. *Shape complementarity at protein/protein interfaces.*
J Mol Biol. **1993** Dec 20;234(4):946-50.

[101] Chen R, Mintseris J, Janin J, Weng Z. *A protein-protein docking benchmark.*
Proteins. **2003** Jul 1;52(1):88-91.

[102] Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. *Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences.*
Crit Rev Biochem Mol Biol. **1996** Apr;31(2):127-52.

[103] Lijnzaad P, Argos P. *Hydrophobic patches on protein subunit interfaces: characteristics and prediction.*
Proteins. **1997** Jul;28(3):333-43.

[104] Zhou HX, Shan Y. *Prediction of protein interaction sites from sequence profile and residue neighbor list.*
Proteins. **2001** Aug 15;44(3):336-43.

[105] Korn AP, Burnett RM. *Distribution and complementarity of hydropathy in multisubunit proteins.*
Proteins. **1991**;9(1):37-55.

[106] Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. *Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect.*
Protein Sci. **1997** Jan;6(1):53-64.

[107] Larsen TA, Olson AJ, Goodsell DS. *Morphology of protein-protein interfaces.*
Structure. **1998** Apr 15;6(4):421-7.

[108] Vakser IA, Aflalo C. *Hydrophobic docking: a proposed enhancement to molecular recognition techniques.*
Proteins. **1994** Dec;20(4):320-9.

[109] Young L, Jernigan RL, Covell DG. *A role for surface hydrophobicity in protein-protein recognition.*
Protein Sci. **1994** May;3(5):717-29.

[110] Bogan AA, Thorn KS. *Anatomy of hot spots in protein interfaces.*
J Mol Biol. **1998** Jul 3;280(1):1-9.

[111] Source: **http://www.expasy.org/tools/pscale/A.A.Swiss-Prot.html**

[112] Jackson RM. *Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem.*
Protein Sci. **1999** Mar;8(3):603-13.

[113] Lange C, Hunte C. *Crystal structure of the yeast cytochrome bc1 complex with its bound substrate cytochrome c.*
Proc Natl Acad Sci U S A. **2002** Mar 5;99(5):2800-5.

[114] Cummings MD, Hart TN, Read RJ. *Atomic solvation parameters in the analysis of protein-protein docking results.*
Protein Sci. **1995** Oct;4(10):2087-99.

[115] Wallqvist A, Covell DG. *Docking enzyme-inhibitor complexes using a preference-based free-energy surface.*
Proteins. **1996** Aug;25(4):403-19.

[116] Jackson RM, Sternberg MJ. *A continuum model for protein-protein interactions: application to the docking problem.*
J Mol Biol. **1995** Jul 7;250(2):258-75.

[117] Weng Z, Vajda S, Delisi C. *Prediction of protein complexes using empirical free energy functions.*
Protein Sci. **1996** Apr;5(4):614-26.

[118] Jackson RM, Gabb HA, Sternberg MJ. *Rapid refinement of protein interfaces incorporating solvation: application to the docking problem.*
J Mol Biol. **1998** Feb 13;276(1):265-85.

[119] Dobson CM. Protein conformation. *Hinge-bending and folding.*
Nature. **1990** Nov 15;348(6298):198-9.

[120] Sandak B, Wolfson HJ, Nussinov R. *Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers.*
Proteins. **1998** Aug 1;32(2):159-74.

[121] May A, Zacharias M. *Accounting for global protein deformability during protein-protein and protein-ligand docking.*
Biochim Biophys Acta. **2005** Dec 30;1754(1-2):225-31. Epub 2005 Sep 12.

[122] Hoppe C, Schomburg D. *Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential.*
Protein Sci. **2005** Oct;14(10):2682-92. Epub 2005 Sep 9.

[123] Heinig M, Frishman D. *STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins.*
Nucleic Acids Res. **2004** Jul 1;32(Web Server issue):W500-2.

[124] Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. *Protein-Protein Docking Benchmark 2.0: an update.*
Proteins. **2005** Aug 1;60(2):214-6.

[125] Huang B, Schröder M. *Using residue propensities and tightness of fit to improve rigid-body protein-protein docking.*
GCB **2005** Proceeding Page 159-173.

[126] Kumar S, Tamura K, Nei M. *MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.*
Brief Bioinform. **2004** Jun;5(2):150-63.

[127] Bradford JR, Westhead DR. *Improved prediction of protein-protein binding sites using a support vector machines approach.*
Bioinformatics. **2005** Apr 15;21(8):1487-94. Epub 2004 Dec 21.

[128] Neuvirth H, Raz R, Schreiber G. *ProMate: a structure based prediction program to identify the location of protein-protein binding sites.*
J Mol Biol. **2004** Apr 16;338(1):181-99.

[129] Lo SL, Cai CZ, Chen YZ, Chung MC. *Effect of training datasets on support vector machine prediction of protein-protein interactions.* *Proteomics.* **2005** Mar;5(4):876-84.

[130] Song J, Tang H. *Support vector machines for classification of homo-oligomeric proteins by incorporating subsequence distributions.* *Journal of Molecular Structure-Theochem* **2005** 722(1-3):97-101.

[131] Yan C, Honavar V, Dobbs D. *Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach.* *Neural Computing & Applications* **2004** 13(2):123-129.

[132] Winter C, Henschel A, Kim WK, Schroeder M. *SCOPPI: a structural classification of protein-protein interfaces.* *Nucleic Acids Res.* **2006** Jan 1;34(Database issue):D310-4.

[133] Teyra J, Doms A, Schroeder M, Pisabarro MT. *SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces.* *BMC Bioinformatics.* **2006** Mar 2;7:104.

[134] Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. *BIND--The Biomolecular Interaction Network Database.* *Nucleic Acids Res.* **2001** Jan 1;29(1):242-5.

[135] Mintseris J, Weng Z. *Structure, function, and evolution of transient and obligate protein-protein interactions.* *Proc Natl Acad Sci U S A.* **2005** Aug 2;102(31):10930-5. Epub 2005 Jul 25.

[136] De S, Krishnadev O, Srinivasan N, Rekha N. *Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different.* *BMC Struct Biol.* **2005** Aug 16;5:15.

[137] Block P, Paern J, Hullermeier E, Sanschagrin P, Sotriffer CA, Klebe G. *Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms.* *Proteins.* **2006** Nov 15;65(3):607-22.

[138] Rodier F, Bahadur RP, Chakrabarti P, Janin J. *Hydration of protein-protein interfaces.* *Proteins.* **2005** Jul 1;60(1):36-45.

[139] Chakrabarti P, Janin J. *Dissecting protein-protein recognition sites.* *Proteins.* **2002** May 15;47(3):334-43.

[140] Bahadur RP, Chakrabarti P, Rodier F, Janin J. *Dissecting subunit interfaces in homodimeric proteins.*
Proteins. **2003** Nov 15;53(3):708-19.

[141] Dill KA. *The meaning of hydrophobicity.*
Science. **1990** Oct 12;250(4978):297-8.

[142] Valdar WS, Thornton JM. *Protein-protein interfaces: analysis of amino acid conservation in homodimers.*
Proteins. **2001** Jan 1;42(1):108-24.

[143] Liddington RC. *Structural basis of protein-protein interactions.*
Methods Mol Biol. **2004**;261:3-14.

[144] Schoelkopf B, Smola AJ. *Learning with kernels.*
Cambridge, MA: MIT Press. **2002** p. 644.

[145] Kottha S, Schroeder M. *Classifiying permanent and transient protein interactions.*
GCB **2006**; Proceedings:54-63.

[146] R Development Core Team. R: *A Language and Environment for Statistical Computing. [http://www.r-project.org].*
R Foundation for Statistical Computing, Vienna, Austria [ISBN 3-900051-07-0] **2005.**

[147] Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. *e1071: Misc functions of the department of statistic (e1071).*
TU Wien. R package version 1.5-8 **2005**.

[148] Chang C, Lin C. *LIBSVM: a Library for Support Vector Machines.*
[http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf]. ***2005***.

Acknowledgements

Initially, I would like to dedicate my thanks to my supervisor Prof. Dr. Volkhard Helms. He provided me with new motivating ideas in moments of frustration. Thanks also to Prof. Dr. Hans-Peter Lenhof for the interest to my thesis. As stated in the corresponding chapters, I'd like to say thanks to Peter Walter for constructing a very powerful database, Kerstin Kunz for modifying and implementing further scoring functions to BDOCK, and Bingding Huang for implementing FFT docking using the BALL library. I thank Yungki Park, Barbara Hutter, and Susanne Eyrisch for revising my thesis. Next, I owe my room colleague Dr. Michael Hutter thanks for being a very helpful roommate over the past years. For most of the technical support and assistance I would like to thank Dr. Tihamér Geyer. Thanks also to my colleagues and coworkers Wei Gu, Dr. Rainer Böckmann, Kerstin Gronow-Pudelek, and former colleagues Dr. Alexander Spaar, Saurabh Kumar Shakya, Dr. Christian Gorba, and Dr. Tomaso Frigato for having a great time on and outside of work.

Last but not least I thank my wife Andreja Ansari, my parents Shahla and Nosrat Ansari, and my brother Nariman Ansari for giving me the strength to go through a number of tough moments and sleepless working nights.