Efficient comprehensive scoring of docked protein complexes - a machine learning approach

Inaugural - Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Oliver Sven Martin
aus Kaiserslautern

Köln 2006

Berichterstatter:	Prof. Dr. D. Schomburg
	Prof. Dr. R. Schrader
Tag der mündlichen Prüfung:	12. Juni 2006

Meinen Eltern

SCIENCE has *explained* nothing; the more we know the more fantastic the world becomes and the profounder the surrounding darkness.

Aldous Huxley, 1894-1963

Acknowledgements

Writing a decent acknowledgement has been among the hardest parts of my thesis work finding the task for a perfect acknowledgement to be a crucial question. Whom to acknowledge at which point in which order and in how much detail? Too many acknowledgements are preferable to too few, but who wants to read page after page of acknowledgements? What about those people who feel aggrieved after reading this acknowledgement and not finding their name listed here or listed in the wrong order? Instead of only thanking those people which I commemorate easiest due to physical or temporal vicinity, I would rather like to thank generically all those who contributed to this work. All those who helped to built the fundamentals for this work, who contributed to the development of new ideas, who motivated and inspired me and who offered their support and friendship; all those forgotten! May each of you find your own name in this acknowledgement in the order you think appropriate. Be assured that I feel obliged to you and have thought long about you and the appropriate way of expressing my gratitude to you.

- My sincere thanks -

Abstract

Biological systems and processes rely on a complex network of molecular interactions. The association of biological macromolecules is a fundamental biochemical phenomenon and an unsolved theoretical problem crucial for the understanding of complex living systems. The term protein-protein docking describes the computational prediction of the assembly of protein complexes from the individual subunits. Docking algorithms generally produce a large number of putative protein complexes. In most cases, some of these conformations resemble the native complex structure within an acceptable degree of structural similarity. A major challenge in the field of docking is to extract the near-native structure(s) out of this considerably large pool of solutions, the so called scoring or ranking problem. It has been the aim of this work to develop methods for the efficient and accurate detection of near-native conformations in the scoring or ranking process of docked protein-protein complexes. A series of structural, chemical, biological and physical properties are used in this work to score docked protein-protein complexes. These properties include specialised energy functions, evolutionary relationship, class specific residue interface propensities, gap volume, buried surface area, empiric pair potentials on residue and atom level as well as measures for the tightness of fit. Efficient comprehensive scoring functions have been developed using probabilistic Support Vector Machines in combination with this array of properties on the largest currently available protein-protein docking benchmark. The established scoring functions are shown to be specific for certain types of protein-protein complexes and are able to detect near-native complex conformations from large sets of decoys with high sensitivity. The specific complex classes are Enzyme-Inhibitor/Substrate complexes, Antibody-Antigen complexes and a third class denoted as "Other" complexes which holds all test cases not belonging to either of the two previous classes. The three complex class specific scoring functions were tested on the docking results of 99 complexes in their unbound form for the above mentioned categories. Defining success as scoring a 'true' result with a *p-value* of better than 0.1, the scoring schemes were found to be successful in 93%, 78% and 63% of the examined cases, respectively. The ranking of near-native structures can be drastically improved, leading to a significant enrichment of near-native complex conformations in the top ranks. It could be shown that the developed scoring schemes outperform five other previously published scoring functions.

Zusammenfassung

Biologische Systeme beruhen auf komplexen Netzwerken molekularer Interaktionen. Die Interaktion biologischer Makromoleküle stellt ein fundamentales biochemisches Phänomen dar, sowie ein ungelöstes theoretisches Problem von herausragender Bedeutung für das Verständnis komplexer lebender Systeme. Als *Protein-Protein Docking* wird die computergestütze Vorhersage der Assoziation von Proteinkomplexen aus den individuellen Untereinheiten bezeichnet. Dockingalgorithmen produzieren im Allgemeinen eine sehr hohe Anzahl hypothetischer Komplexanordnungen, von denen meist nur einige wenige der korrekten, nativen Lösung ähnlich sind. Eine der grossen Herausforderungen im Bereich des Dockings besteht im Herausfiltern der wenigen nahe-nativen Strukturen aus der grossen Menge von Lösungsvorschlägen. Dieses wird auch als *Scoring-* oder *Rankingproblem* bezeichnet. Ziel dieser Arbeit war es, Methoden zur effizienten und akkuraten Detektion von nahe-nativen Lösungen während der Bewertungsphase von gedockten Proteinkomplexen zu entwickeln.

Eine Reihe von strukturellen, chemischen, biologischen und physikalischen Parametern wurde verwendet, um Komplexanordungen, wie sie als Lösungsvorschläge eines Dockingalgorithmus enstehen, zu bewerten. Diese Bewertungsschemata beinhalten spezialisierte Energiefunktionen molekularer Fragmente, evolutionäre Verwandtschaft, komplexklassenspezifische Wahrscheinlichkeitsverteilungen von Residuen, Lückenvolumen, die Grösse der verborgenen Oberfläche, emprische Paarpotentiale auf atomarer und Aminosäurebene sowie ein Mass für die Festigkeit der Bindung. Unter Verwendung des derzeit grössten Datensatzes von Protein-Protein Docking Testfällen wurden Verfahren des überwachten maschinellen Lernens in Form von probabilistischen *Support Vector Machines* trainiert,s um umfassende effiziente Bewertungsfunktionen für drei spezifische Klassen von Proteinkomplexen zu erstellen. Bei diesen Dockingklassen handelt es sich um Enzym-Inhibitor bzw. Enzym-Substrat und Antikörper-Antigen Komplexe sowie eine dritte Klasse, der alle weiteren Testfälle zugordnet werden, die keiner der beiden bisherigen Kategorien angehören. Die entwickelten Bewertungsfunktionen sind hochspezifisch für die einzelnen Kategorien von Proteinkomplexen und in der Lage, nahe-native Lösungen mit hoher Sensitivität aus einer grossen Anzahl potentieller Komplexanordnungen heraus zu erkennen. Eine Sortierung der

Lösungsvorschläge durch Anwendung der Bewertungsfunktionen führt zu einer signifikanten Anreicherung von nahe-nativen Komplexen in den oberen Rängen. Die drei entwickelten spezifischen Bewertungsfunktionen wurden an Dockingergebnissen für 99 Testfälle erprobt, bei denen versucht wird, native Komplexe aus den ungebunden Strukturen der einzelnen Untereinheiten vorherzusagen. Definiert man ein "korrektes" Ergebnis über einen Wahrscheinlichkeitswert (*p-value*) von 0,1 oder besser, so sind die entwickelten Bewertungsfunktionen in 93%, 78% und 63% der untersuchten Fälle erfolgreich. Ein Vergleich mit fünf publizierten Bewertungsfunktionen für Protein-Protein Docking zeigt, dass die komplexklassenspezifischen Bewertungsfunktionen den jeweils einzelnen Methoden in der Anwendung überlegen sind.

Definitions and abbreviations

ASA	accessible surface area
Ångstrøm	1 Å$= 10^{-10}$m
$spec^-$	specifictiy; reliability of false/negative predictions
$spec^+$	specificity; reliability of true/positive predictions
acc	accuracy
ACE	atomic contact energies; an atom-atom pair potential
AUC	area under the curve
avg.	average
avgTF	scoring scheme based on average temperature factor in the interface area
BSV	Bounded Support Vectors
BurSurf	buried surface area; are occluded from solvent by contact surfaces of complexed proteins
CAPRI	Critical Assessment of PRedicted Interactions; academic challenge for blind protein interaction predictions
Cons	protein docking scoring scheme based on amino acid conservation scores
ConsOE	protein docking scoring scheme based on amino acid conservation scores with an over emphasis on residues with a high interface propensity
fn	false negative
fp	false positive
fval	f-value; harmonic average between sensitivity and specificity
GapVol	gap volume; volume inbetween and delimited by contact surface of two or more proteins
geo	score/rank according to geometric fit
IF	improvement factor
mcc	Matthews correlation coefficient
NhcR	number of highly conserved residues
NhvR	number of highly variable residues
P	probability
PairPot	atom-atom pair potential
pred	score/rank according to SVM predictor
red.	reduction
RIP	residue interface propensities
RIP_{AA}	residue interface propensities for Antibody-Antigen complexes
RIP_{EI}	residue interface propensities for Enzyme-Inhibitor/Substrate complexes
RIP_{UNI}	residue interface propensities; universally applicable
RMSD	root mean square deviation; a measure for structural similarity

$RMSD_{iC_\alpha}$	root mean square deviation of interface C-alpha atoms
ROC	Receiver Operator Characteristics; plot of true positive against false positive rate
Rpscore	an empiric residue-residue pair potential
SE	solvent effect
sens	sensitivity
SV	Support Vectors
SVM	Support Vector Machines; a machine learning method
SVM_{AbAg}	SVM-based scoring scheme developed for Antibody-Antigen complexes
SVM_{EI}	SVM-based scoring scheme developed for Enzyme-Inhibitor/Substrate cimplexes
SVM_{Oth}	SVM-based scoring scheme developed for complexes of type "Other" (non-Antibody-Antigen and non-Enzyme-Inhibitor/Substrate complexes)
tn	true negatve
ToF	Tightness of Fit; a scoring scheme for docked protein-complexes
ToF_{UNI}	Tightness of Fit; universally applicable
ToF_{EI}	Tightness of Fit; specialised for Enzyme-Inhibitor/Substrate complexes
ToF_{AA}	Tightness of Fit; specialised for Antibody-Antigen complexes
tp	true positive

Aminoacids

Alanine	ALA	A
Cysteine	CYS	C
Aspartate	ASP	D
Glutamate	GLU	E
Phenylalanine	PHE	F
Glycine	GLY	G
Histidine	HIS	H
Isoleucine	ILE	I
Lysine	LYS	K
Leucine	LEU	L
Methionine	MET	M
Asparagine	ASN	N
Proline	PRO	P
Glutamine	GLN	Q
Arginine	ARG	R
Serine	SER	S
Threonine	THR	T
Valine	VAL	V
Tryptophane	TRP	W
Tyrosine	TYR	Y

List of Tables

List of Figures

Contents

1 Introduction

"He has half the deed done who has made a beginning."

Horace, 65-8 B.C.

According to the conventional definition of life, an organism in question must exhibit the following five stages of a living system at least once during their existence: growth, motion, reproduction, metabolism, and response to stimuli. These parameters alone, however, may be inadequate for proper classification without further specification. For example, a mule is a living system, yet it cannot reproduce. Conversely, a non-living entity such as fire may experience all five stages on some level. Biochemistry focuses specifically on the aspect of metabolism to define a living system and implies that the energy gained through metabolism is utilised to maintain the living state by flowing into a coordinated regulatory network of molecular interactions. This network is the fundamental basis for reactivity and all the other phenomena used in the conventional definition of a living system. Implicitly, a huge quantity of the ongoing processes in every living organism are based on, regulated or mediated by molecular recognition mechanisms, thus the activity of a living cell can be portrayed as a network of interactions. Such an interaction network could never be coordinated without a high level of specificity. The specificity is mostly provided by the enormous structural and physico-chemical variability of biological macromolecules like proteins and nucleic acids, that are involved in the transfer of biological information.

Protein-protein interactions play a significant role in these processes for example in signal cascades or gene regulation. In the proteomics era, where experimental high-throughput methods like e.g. the yeast two-hybrid system yield growing amounts of putative protein interaction data, the large quantity of data can no longer be handled by experimental methods alone. Instead it requires the computer aided simulation methods of bioinformatics to complement this knowledge. The exploration, understanding and detailed knowledge of complete protein interaction networks can only be achieved by combining the often time consuming experimental methods like structure solution by X-ray crystallography or NMR spectroscopy with the data management

facilities and theoretical predictions provided by bioinformatics. Predictive methods for protein-protein interactions are of special interest and importance where experimental methods fail, e.g. for such short-term transient interactions which are not accessible by the mentioned experimental methods due to their low stability and short half-life (Eisenstein and Katchalski-Katzir, 2004).

1.1 Protein-protein assemblies

Protein-protein interactions play diverse roles in biology and differ based on the composition, affinity and half-life of the association. In vivo, the localisation, concentration and local environment of a protomer (subunit of an oligomeric protein complex) can affect the interaction between protein domains and are vital to control the composition and oligomeric state of protein complexes. Since a change in quaternary structure is often coupled with biological function or activity, transient protein-protein interactions are important biological regulators.

1.1.1 Types of protein-protein interactions

Protein-protein interactions are often categorised into distinct types according to their composition, *in vivo* stability and lifetime (Nooren and Thornton, 2003a):

- Homo- and hetero-oligomeric complexes

  Protein-protein interactions occur between identical or non-identical chains (i.e. homo- or hetero-oligomers). Oligomers of identical or homologous protomers can be organised in an isologous or heterologous way. An isologous association involves the same surface on two monomers forming an interface with matching surfaces (e.g. Arc repressor and lysin; Figure 1.1 (a) and (c)), related by a 2-fold symmetry axis. In contrast to an isologous association that can only further oligomerise using a different interface (e.g. form a dimer of dimers with three 2-fold axes of symmetry), heterologous assemblies use different interfaces that, without a closed (cyclic) symmetry, can lead to infinite aggregation (cf. Figure 1.2 (a,b)).

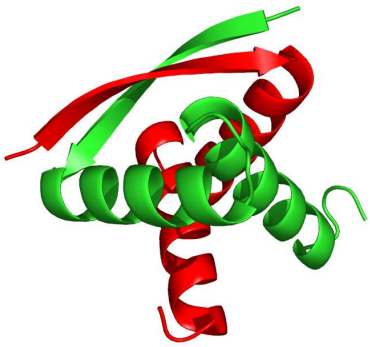
- Non-obligate and obligate complexes

  As well as composition, two different types of protein-protein complexes can be distinguished on the basis of whether a complex is obligate or non-obligate. In an obligate protein-protein interaction, the protomers are not found as stable structures on their own in vivo.
  Such complexes are generally also functionally obligate; for example, the Arc repressor dimer (Figure 1.1 (a)) is essential for DNA binding. Many of the hetero-oligomeric structures in the PROTEIN DATA BANK (PDB) (Berman et al., 2000) involve non-obligate interactions of protomers that exist independently, such as intracellular signaling complexes and antibody-antigen, receptor-ligand and enzyme-inhibitor (e.g. thrombin-rhodniin; Figure 1.1 (e)) complexes. The components of such protein-protein complexes are often initially not co-localised and thus need to be independently stable. However, some homo-oligomers, which by definition are co-localised, can also form non-obligate assemblies (e.g. sperm lysin; Figure 1.1 (c)).

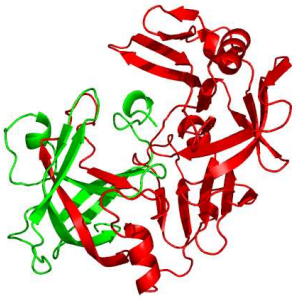
- Transient and permanent complexes

  Protein-protein interactions can also be distinguished based on the lifetime of the complex. In contrast to a permanent interaction that is usually very stable and thus only exists in its complexed form, a transient interaction associates and dissociates in vivo. One can distinguish between weak transient interactions that feature a dynamic oligomeric equilibrium in solution, where the interaction is broken and formed continuously (e.g. lysin; Figure 1.1 (c)), and strong transient associations that require a molecular trigger to shift the oligomeric equilibrium. Structurally or functionally obligate interactions are usually permanent, whereas non-obligate interactions may be transient or permanent.

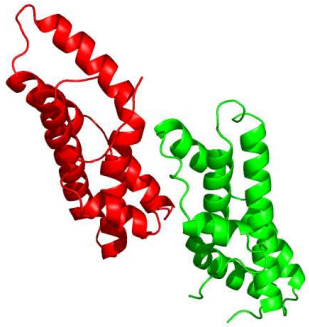
It is important to note that many protein-protein interactions cannot be classified according to such unique distinct types. Rather, a continuum exists between non-obligate and obligate interactions (Nooren and Thornton, 2003b), and the stability of all complexes is highly dependant on the physiological conditions and environment. An interaction may be mainly transient in vivo but become permanent under certain



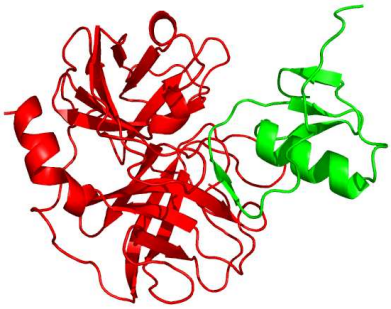
(a) 1myk



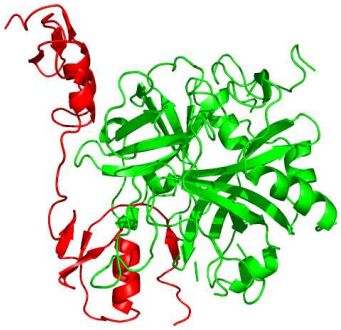
(b) 1lyw



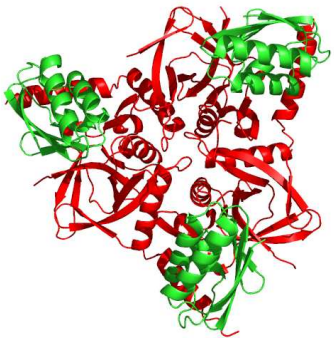
(c) 1lyn



(d) 1ppf

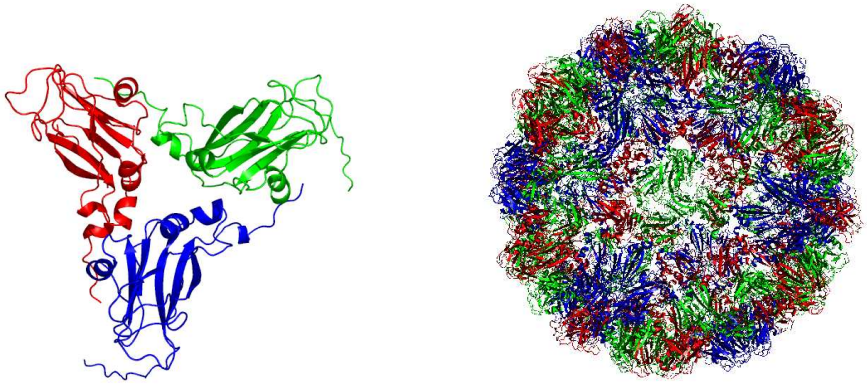


(e) 1tbr



(f) 1kkl

Figure 1.1: Selected examples for protein-protein interaction types (4-letter PDB-identifier given): (a) obligate homomeric complex: P22 ARC repressor, (b) obligate heteromeric complex: human cathepsin D, (c) non-obligate homomeric complex: sperm lysin, (d) non-obligate heteromeric complex: human leukocyte elastase / turkey ovomucoid inhibitor, (e) non-obligate permanent heteromeric complex: thrombin / rhodniin inhibitor, (f) non-obligate transient heteromeric complex: L. casei protein kinase Hprk / B. subtilis Hpr (obligate permanent interaction in Hprk trimer (red), transient binding to Hpr (green)).



(a) 1cwp **(b)** 1cwp-assumed biological unit

Figure 1.2: Selected examples for protein-protein interaction types (4-letter PDB-identifier given): (a) trimeric unit of viral coat protein: cowpea chlorotic mottle virus (b) assumed biological unit of viral coat protein: spherical virus capsid (consisiting of 60 trimeric units as depicted in (a), identical colour coding)

cellular conditions (Nooren and Thornton, 2003a). Folding data, as well as data on the dynamics of the assembly at different physiological conditions or environments, are often not available. However, the sub cellular localisation of subunits and the function of the protein will often suggest the biologically relevant type of interaction. For example, interactions in intracellular signaling are expected to be transient, since their function requires a ready association (Rittinger et al., 1997b,a). Ultimately, all interactions and complexation processes are driven by the concentration of the components and the free energy of the complex relative to alternate states.

1.1.2 Specificity of protein-protein interactions

The specificity of protein-protein interaction is composed of two major factors: the possibility of forming a more or less stable binding to its predestined binding partner(s) and the potentially lower possibility of association to other protomers in an equally stable and favorable way. It is well known that the binding specificity of protein-protein interactions is mostly accomplished by relatively small structural changes in the contact area of the binding partners - the so called interface region - rather than spacious structural rearrangements (Sear, 2004). A single point mutation leading to

an amino acid change in the interface region of any of the binding partners can cause a complete loss or change of binding specificity as well as create an entirely new one. The consequences of such a change in binding specificity can be quite drastic as the following prominent examples of human hereditary diseases show:

(i) An example for a reduction or loss in binding specificity due to a single amino acid change is *osteogenesis imperfecta*, commonly known as the "brittle bone" disorder, a genetic disorder characterised by bones that break easily, often from little or no apparent cause. The molecular cause for this disease is a defective collagen assembly (Vogel et al., 1987). Collagen is a family of related structural proteins which are vital to the integrity of many tissues including skin and bones. The mature collagen molecule is comprised of three peptide chains wound in a triple helix. In order to form the triple helix, collagen peptide chains have a special repeating structure consisting of a specific three amino acid pattern. A point mutation which, by changing a single amino acid, disrupts that pattern, will either disturb the association of chains or prevent the triple helix formation and may have very severe consequences. One mutant chain can disrupt a triple helix with two wild type chains, effectively disturbing the functional unit in its stabilising efficacy.

(ii) The single substitution of valine for glutamic acid at position six of the beta-globin polypeptide chain in human haemoglobin gives rise to *sickle cell anaemia* in homozygote individuals. The modified chain reveals an extended binding specificity to itself and therefore develop a tendency to crystallise at low oxygen concentrations, forming threads of haemoglobin molecules which in turn evoke the sickle-like shape of erythrocytes that gave the disease its name (Rodgers, 1997).

Often, proteins are only biologically active in the complexed oligomeric state. A loss of the ability to form the relevant oligomer therefore can lead to a loss in biological activity, as it has been described by Bennett et al. (1994) for a single amino acid substitution in dimeric proteins.
Single mutations on primary structure level that lead to a loss of the binding affinity of transcription factors can even lead to gene knockout (Rausa et al., 2004). The ability of introducing an entirely new binding specificity by a single point mutation without

taking the intermediate stage of a non-specific interface is a critical pre-requisite for the evolution of interfaces (Xu et al., 1998) (see section 1.1.3).

1.1.3 Evolution of protein-protein interactions

The structure and affnity of a protein-protein interaction is tuned to its biological function as well as the physiological environment and control mechanism. Protein-protein interactions presumably evolve to optimise *functional efficacy*. Not necessarily are strong interactions involved, since weak transient interactions that are efficiently controlled are of similar importance in cellular processes. Obligate complexes may simply reflect the need for stability or the evolution of a function that requires both protomers. For example, symmetric DNA-binding modules, designed to bind to an equally symmetric macromolecule, or inter-subunit active sites with catalytic residues on different subunits, that would be inactive as separate proteins. While some oligomerisations are obligate from a functional perspective, others may seem incidental to function (e.g. oligomerisation of cytokines whose primary function lies in receptor binding as a monomer). It might seem that such an interaction survives because there is simply no selective pressure to reject it from the evolutionary path, but on the other hand redundancy is also an evolutionary principle, providing a backup in the case of malfunctions of the first instance. The evolution of a protein-protein interaction may also be related to folding, especially in the case of obligate complexes, where folding of the individual protomers and oligomerisation occur concurrently (Xu et al., 1998). In contrast, in non-obligate interactions, each protomer folds independently and the interaction site has presumably evolved on the surface of the stable monomer. Some oligomers may evolve through domain swapping that involves a rearrangement of domains where inter-domain interactions are replaced by inter-monomer interactions (Bennett et al., 1994). Varying oligomeric states or structures within a homologous protein family can give further hints on the evolution of the family. For a conserved oligomeric state, the residues at the interface are preferentially conserved compared with the rest of the surface (Valdar and Thornton, 2001). However, in large families that have members with varying oligomeric states or structures, these residues are found to be less conserved, as expected (Nooren and Thornton, 2003b).

1.2 Protein-protein interfaces

Many studies have been accomplished in order to gain knowledge on the general nature of the interacting surface areas involved in protein-protein interactions, the so called protein-protein interfaces (Bogan and Thorn, 1998; Larsen et al., 1998; Jones and Thornton, 1996). The individual aspect of such studies, ranging from the size and shape of interfaces to their composition and physico-chemical properties will be discussed in the following subsections.

1.2.1 Structural characteristics of protein-protein interfaces

The interface of a - notional binary - protein complex depicts those parts of the surface area of the protomers where they are in close contact to each other. Those contacting sections of the protomers' surface are often denoted as interface patches.

1.2.1.1 Geometric properties of interface patches

Areas on the surface of the individual protomers of a protein-protein complex which are in close proximity to each other and presumably involved in establishing the interaction are denoted as interface areas. The interface areas consist of one or more areal contiguous fractions of the proteins' surfaces denoted as interface or recognition patches. In statistical surveys, Janin and Chothia (1990) list the average interface area with 1600 ± 400 Å^2 giving an estimate for the standard interface size. This is equivalent to 170 ± 39 surface atoms, or 85 atoms per recognition patch. Often, but not necessarily, a protein-protein interface constitutes of a single interface patch. Chakrabarti and Janin (2002) refined the statistical analysis and showed that multi patch interfaces can be considerably larger than single patch interfaces but generally contain at least one pair of patches that is equivalent in size to a single patch interface. While Jones and Thornton (1997) have noted, that protein-protein interfaces tend to be planar, Chakrabarti and Janin (2002) found out, that some are clearly non-planar.

1.2.1.2 Physico-chemical properties of interface patches

By screening a large number of alanine mutants for which the change in free energy of binding upon mutation to alanine has been measured, Bogan and Thorn (1998)

discovered that the free energy of binding is not evenly distributed across interfaces. Instead, there are so called hot spots of binding energy, made up of a small subset of residues in the dimer interface. These hot spots are enriched in hydrophobic amino acids, and are surrounded by energetically less important residues that most likely serve to occlude bulk solvent from the hot spot. Occlusion of solvent was found to be a necessary condition for highly energetic interactions.
Lo Conte et al. (1999) as well as Chakrabarti and Janin (2002) consequently distinguished two regions of interface patches: the core region, which constitues of those atoms that are solvent accessible in the unbound state and will loose their contact to the solvent within the transition to the complexed state, and the rim region, which is composed of those atoms that remain at least partially solvent accessible in the complexed state. The idea behind such a distinction is that the rim region of a recognition patch acts like a sealing, shielding the core region of the interface from the solvent and thus enabling a drastic change in the medium that transmits interaction forces.
Further studies showed, that energetic hot spots correlate well with sequentially highly conserved residues (Hu et al., 2000; Ma et al., 2001, 2003; Halperin et al., 2004). If it is only a limited number of residues that make up for most of the binding free energy of a functional protein-protein complex and if protein-protein interactions evolve to optimise functional efficacy, these residues should clearly be conserved during evolution. Lijnzaad and Argos (1997) found in their studies, that hydrophobicity plays an important role in complex formation by detecting hydrophobic surface areas of the protomers (Lijnzaad et al., 1996) and a subsequent statistical analysis of these patches in interface and non-interface regions of a set of protein-protein complexes. In 90% of the cases, the largest or second largest hydrophobic surface patch was overlapping with the interface region. The fraction and distribution of hydrophobic patches vary significantly with the type of protein-protein complex. Large hydrophobic contact areas are predominant mostly in homomeric obligate permanent complexes. This gives also an explanation for their permanent and obligate binding: Exposing such large hydrophobic surface areas directly to the aqueous environment of a living cell would destabilise the protomers beyond means.

1.2.1.3 Composition of interface patches

Since interface regions of protein-protein complexes seemingly differ from the rest of the protein surface in their physico-chemical properties, investigations aiming for the detection of a difference in the composition of their primary building units - the amino acids - lie at hand. Jones and Thornton (1997) and Lo Conte et al. (1999) calculated propensities for each of the 20 proteinogen amino acids to be part of an interface region on different datasets of protein-protein complexes and came to similar results: The by far highest interface propensities were assigned to the large hydrophobic amino acids Tryptophane and Tyrosine, followed by Methionin and Phenylalanine. Cysteine obtained a rather high value as well due to its ability to form highly stabilising disulfide bridges, followed by Histidine, Isoleucine and Leucine. A propensity around zero was assigned to the smallest amino acid Glycine, while the highly hydrophilic amino acids Asparagine and Aspartate, Glutamine and Glutamate, Lysine, Proline, Serine and Threonine were assigned negative values, clearly being least abundant in interface areas. Chakrabarti and Janin (2002) were able to further refine these interface propensities by splitting the interface patches into a core and rim region, clearly indicating that the more hydrophilic an amino acid is, the higher the difference in propensity between the core and the rim region becomes. This fortifies the theory of a hydrophobic core surrounded by a slightly more hydrophilic rim on interface patches.

1.2.1.4 Native interfaces versus crystal contacts

While native protein-protein interfaces are highly specific, thus making complex formation a directed process, non-native interfaces, as they occur e.g. in protein crystals, are often randomly induced. This is among others caused by the fact, that proteins are dynamic structures and that the concentration of protein in a crystal is significantly higher than under natural conditions. In order to distinguish native interfaces from crystal-packing contacts, geometric and physical chemical properties can be consulted as distinction criteria:
The size of random interfaces is consistent with those of native interfaces, as an analysis of non-native random protein-protein associations generated by computer aided simulations yielded (Janin and Rodier, 1995). Even though the overall interface area of non-specific interfaces does not distinguish them from functional contact surfaces, the

size of the individual interface patches can be used to discriminate highly fragmented crystal contact areas (45% of which show sizes of less than 100 Å^2; Carugo and Argos (1997)) from functional interfaces. Furthermore, non-specific interfaces are found to be less compact in terms of atomic packing. The chemical and amino acid compositions of large crystal-packing interfaces resemble the protein solvent-accessible surface. These interfaces are less hydrophobic than in homodimers and contain much fewer fully buried atoms. Using a residue propensity score and a hydrophobic interaction score to assess preferences seen in the chemical and amino acid compositions of the three different types of interfaces, as well as indexes to evaluate the atomic packing, Bahadur et al. (2004) were able to distinguish crystal contacts from native protein-protein interfaces with accuracies up to 95%.

1.3 Characterisation of protein-protein interactions

With the amount of data available on genetic interactions, a lot of attention has been drawn on systems biology, in particular biomolecular interactions. Even though a large number of methods has been developed to detect, examine, predict and quantify protein-protein interactions, it is still not possible to determine the full interaction network of a complete cell. These methods differ with respect to their aim as well as the nature and the amount of details their results provide. The most common goals are the determination of binding partners, the determination of the complex structure, the quantification of the binding force, and the examination of binding kinetics.

1.3.1 Experimental methods

Biochemical and biophysical experiments are widely used to gain insight into biomolecular interactions. The following section gives a brief overview on the basic working principles and the information gained by selected experimental methods. The methods can be classified according to the detail level of information they provide about protein complexes which reach from the determination of individual binding residue pairs to the complete determination of complex structures in atomic detail and the quantification of the binding force.

1.3.1.1 Determination of binding partners, binding regions or interaction restraints

- Protein affinity chromatography

  A protein can be covalently coupled to a matrix such as Sepharose under controlled conditions and be used to select ligand proteins that bind and are retained from an appropriate extract. A particular clever and useful variety of this method is the so called Tandem Affinity Purification (TAP) (Rigaut et al., 1999), where, via systematic Polymerase Chain Reaction (PCR) mutagenesis at the 3'-end of the gene, the protein is provided with a specific tag. This tag allows for the purification using an appropriate adapted affinity column. Using the right conditions for eluations makes the purification of complex partners of the tagged protein(s) possible (Gavin et al., 2002), implicitly providing the information which binding partners are involved.

- Affinity Blotting

  In a procedure analogous to the use of affinity columns, proteins can be fractionated by Polyacrylamid Gel Electrophoresis (PAGE), transferred to a nitrocellulose membrane, and identified by their ability to bind a protein, peptide, or other ligand (Vasilescu et al., 2004).

- Immunoprecipitation

  Co-immunoprecipitation is a classical method of detecting protein-protein interactions and has been used frequently in experiments. For this method cell lysates are generated, antibody is added, the antigen is precipitated and washed, and bound proteins are eluted and analysed (Masters, 2004).

- Chemical cross-linking

  The procedure of chemical cross-linking involves three steps. First, the complex (presumably of units P and P') is reacted with a cleavable bi-functional reagent containig a disulfide bridge of the form RSSR', and the R and R' groups react with susceptible amino acid side chains in the protein complex PP'. This reaction forms adducts of the form P-RSSR'- P'. Second, the proteins are fractionated on an SDS-gel in the absence of reducing agents. The gel separates the proteins

based on molecular weight, and cross-linked proteins of the form P-RSSR'-P' migrate as species of greater molecular weight. Third, a second dimension of the SDS-gel is run after treatment of the gel with a reducing agent to cleave the central disulfide bond. Un-cross-linked species align along the diagonal of the 2D-gel, because their molecular weights do not change after reduction. Cross-linked proteins migrate off the diagonal because they migrated as P-RSSR'-P' in the first dimension and as molecules of the form P-RSH and P'-R'SH in the second dimension. The cross-links are identified by their size, which corresponds to that of the un-cross-linked species P and P' (Fancy, 2000).

- Protein probing

  A labeled protein can be used as a probe to screen an expression library in order to identify genes encoding proteins interacting with this probe. Interactions occur on nitrocellulose filters between an immobilised protein and the labeled probe protein. This procedure was automised in large scale on what is known as protein microarrays or proteome chips (Kawahashi et al., 2003).

- Phage display

  Smith (1985) first demonstrated that an E. coli filamentous phage can express a fusion protein bearing a foreign peptide on its surface. These foreign amino acids were accessible to antibody, such that the "fusion phage" could be enriched over ordinary phage by immunoaffinity purification. Smith suggested that libraries of fusion phage might be constructed and screened to identify proteins that bind to a specific antibody. There have been numerous developments in this technology to make it applicable to a variety of protein-protein and protein-peptide interactions.

- Yeast two-hybrid system

  The two-hybrid system (Fields and Song, 1989) is a genetic method that uses transcriptional activity as a measure of protein-protein interaction. It relies on the modular nature of many site-specific transcriptional activators, which consist of a DNA-binding domain and a transcriptional activation domain. The DNA-binding domain serves to target the activator to specific genes that will be expressed, and the activation domain contacts other proteins of the transcriptional machinery to enable transcription. The two-hybrid system is based on the observation

that the two domains of the activator need not be covalently linked and can be brought together by the interaction of any two proteins. The application of this system requires that two hybrids are constructed: a DNA-binding domain fused to protein X, and a transcription activation domain fused to protein Y. These two hybrids are expressed in a cell containing one or more reporter genes. If X and Y interact, they create a functional activator by bringing the activation domain into close proximity with the DNA-binding domain. This can be detected by expression of the reporter genes.

- Mass Spectrometry

  There has been increasing interest in Mass Spectrometry as a tool in structural biology in general, but particularily to obtain information about biomolecular complexes. One approach used is Hydrogen/Deuterium exchange. With this method, the rate of exchange provides information about the accessibility of a residue in question. Rate differences between free and bound forms indicate that a given residue is protected on complex formation and thus probably involved in the interaction (Lanman and Prevelige, 2004). Another possibility is cross-linking, where residues close in space are detected by first covalently linking two molecules by the use of a cross-linking reagent, and then subjecting the resulting material to peptide mass fingerprinting or other protein identification methods (Back et al., 2003).

1.3.1.2 Determination of the complex structure

- X-ray crystallography

  Protein X-ray crystallography provides a detailed picture of the atomic structure of a protein-protein complex. Most of the complex structures known so far have actually been determined by this method which uses the diffraction of X-rays by periodically composed protein crystals. From the resulting diffraction patterns, the relative position of the protein backbone, side chains, down to the individual atoms (depending on the resolution attained) can be calculated. As mono-crystals of the respective protein are needed for this process, this also limitates the method, since especially large and hydrophobic proteins (e.g. membrane complexes) are difficult and often time consuming to crystallise. Another problematic

feature of X-ray diffraction patterns is the fact that they provide a single "frozen" snapshot of the dynamic protein structure in an artificial environment (cf. section 1.2.1.4 on page 10). Though one might argue, that protein crystals can consist of up to 70% of crystal water. The concentration of protein in such a crystal is such comparable to the cytosolic protein concentration (up to 25%).

- Nuclear Magnetic Resonance spectroscopy (NMR)

  NMR spectroscopy relies on the absorption and emission of radio-frequency radiation by the nuclei of certain atoms when they are placed in a magnetic field and facilitates the measurement of inter atomic distances and connectivities. In contrary to the X-ray crystallography, this method allows proteins to be studied in solution, giving full access to the molecules' dynamics via a whole time series of snapshots of the molecule, without the influence of crystal contacts. The method is limitated, due to its complexity, by the size of the molecules for which the relative positions of the atoms can be determined. The "classical" approach, based on the use of intermolecular Nuclear Overhauser Effects (NOE), in combination with Residual Dipolar Couplings (RDC) allows for the determination of protein structures of a sequence length up to a maximum of 300 amino acids. Novel methodologies like Transverse Relaxation Optimised Spectroscopy (TROSY) and Chemical Shift Perturbations (CSP) have alleviated the size limitations for the determination of biomolecular structures in solution up to a mass of 50 kDA (Bonvin et al., 2005). Based on the average mass of the twenty proteinogen amino acids (118.9 DA), this equals an average sequence length of 420 amino acids.

1.3.1.3 Quantification of the binding force

- Isothermal Titration Chromatography (ITC)

  Isothermal Titration Chromatography (ITC) is the most quantitative means available to measure the thermodynamic properties of protein-protein interaction. The procedure is able to determine the stoichiometry of the interaction, the association constant, the free energy, enthalpy, entropy, and heat capacity of binding. ITC measures the binding equilibrium directly by determining the heat

evolved on association of a ligand with its binding partner. In a single experiment, the values of the binding constant, the stoichiometry, and the enthalpy of binding are determined. The free energy and entropy of binding are determined from the association constant. The temperature dependence of the enthalpy parameter, measured by performing the titration at varying temperatures, describes the heat capacity term (Pierce et al., 1999).

- Nanobiotechnology

  The recent progress in nanobiotechnology enabled the direct access to intermolecular forces. One example is the so called Atom Force Microscopy (AFM) which allows to physically measure the absolute binding force between two macromolecules via capillary springs (Clausen-Schaumann et al., 2000). Besides this direct way to measure the binding affinity, there is also the possibility to quantify this force via a comparison of one complex to others posing as a reference, in a procedure known as Congruent Force Intermolecular Test (C-FIT) (Albrecht et al., 2003).

- Surface Plasmon Resonance (SPR)

  This method measures complex formation by monitoring changes in the resonance angle of light impinging on a gold surface as a result of changes in the refractive index of the surface up to 300 nm away. A ligand of interest (peptide or protein in this case) is immobilised on a dextran polymer on the gold coated surface. A protein that interacts with the immobilised ligand is retained on the polymer surface, which alters the resonance angle of impinging light as a result of the change in refractive index brought about by increased amounts of protein near the polymer. Since all proteins have the same refractive index and since there is a linear correlation between resonance angle shift and protein concentration near the surface, this allows to measure changes in protein concentration at the surface due to protein-protein or protein-peptide binding, respectively. Furthermore, the measurements can be done in real time, giving access to the kinetics of the reaction (Malmqvist, 1993).

In the area of functional genomics, the rapidly increasing number of completely annotated genomes accessible reveals the existence of many proteins for which functional

information is incomplete or absent. Especially the methods mentioned above gained on importance as they can be used to screen whole libraries of proteins and are suitable as high-throughput assays. This arises the question of reliability of such methods. The yeast two-hybrid method for example is known to produce a high number of false positive interactions and the assessed liability is about 50% (Sprinzak et al., 2003), which clearly indicates that so far a combination of different methods is needed to uncover the interaction network of a cell. Theoretical methods, which will be attended to in the next section, can further supplement the experimental data.

1.3.2 Theoretical computational methods

Theoretical approaches are used to address the problem of protein-protein interaction prediction. These are classified here according to the information required as input and/or prerequisite.

1.3.2.1 Methods based on genetic information

Computational methods based on genetic information are often used to validate experimental interaction data (e.g. the outcome of yeast two-hybrid experiments) and detect false positive interactions, but can also be used for the prediction of protein function and interaction (Date and Marcotte, 2005).

- Phylogenetic profile comparison

  Phylogenetic profile comparison is based on the pattern of the presence or absence of a given gene in a set of genomes, that is, determining in which organisms the gene is present and in which it is not. Similarity of phylogenetic profiles can be interpreted as being indicative of the functional need for corresponding proteins to be simultaneously present in order to perform a given function in combination. However, although this similarity may suggest a related functional role, a direct physical interaction between the proteins is not necessarily implied (Pellegrini et al., 1999). The main limitations of this approach lie in the fact that it can only be applied to complete genomes, as only then it is possible to rule out the absence of a given gene. Similarly, the method cannot be used with essential proteins that are common to most organisms.

- Conservation of gene neighbourhood

  The organisation of bacterial genomes into regions that tend to code for functionally related proteins, such as operons, is a well-known fact. This neighbourhood relationship becomes even more relevant when it is conserved in different species. The adjacency of genes in various bacterial genomes has been used to predict functional relationships between corresponding proteins (Dandekar et al., 1998). The main limitations of this method is that it is only directly applicable to bacteria, in which the genome order is a relevant property.

- Gene fusion events

  Interactions between proteins can be deduced from the presence in different genomes of the same protein domains, which either form part of a single polypeptide chain (multi-domain protein) or act as independent proteins (single domains). Methods based on recursive sequence searches and multiple sequence alignments have been combined in order to detect such domain fusion events (Marcotte et al., 1999; Enright et al., 1999). By definition, this approach is restricted to shared domains in distinct proteins, a phenomenon whose true extent is still unclear, especially in prokaryotic organisms.

- Similarity of phylogenetic trees

  Based on the assumption that interacting protein pairs coevolve, the corresponding phylogenetic trees of the interacting proteins should show a greater degree of similarity or symmetry than noninteracting proteins would be expected to show. This so called mirrortree method, essentially an extended version of the phylogenetic profile comparison, can be used to identify potentially interacting proteins (Pazos and Valencia, 2001).

1.3.2.2 Sequence based methods

According to Anfinsen's hypothesis that "protein sequence determines structure determines function" (Anfinsen, 1973), many of the properties of a protein can be predicted if only they are known for other proteins of homologous sequence. To a certain amount, this is also true for the identification of protein-protein interfaces, in particular when specialised binding motifs or domains have evolved due to a higher amount of selective

pressure on functional parts of the protein surface which exist in various proteins and communicate certain binding properties. Prominent examples are the Leucine zipper motif or the SH2 and SH3 protein-protein interaction domains. The Leucine zipper is a protein-protein interaction motif in which there is a cyclical occurrence of Leucine residues every seventh residue over short stretches of a protein in an alpha-helix. These Leucine residues project into an adjacent Leucine zipper repeat by interdigitating into the adjacent helix, forming a stable coiled-coil (Landschulz et al., 1988). The SH2 domain has been recognised as a common motif involved in protein-protein interactions in a significant number of proteins. They share a motif of about 100 amino acids that is involved in the recognition of proteins and peptides containing phosphorylated tyrosines. Many proteins have been shown to have an SH3 domain, which varies between about 55 and 75 amino acids in length. Like the SH2 domain, the SH3 domain binds simple peptides with a high degree of sequence specificity and a high affinity. As judged on a qualitative basis, a 10-amino-acid Proline-rich sequence within the domain is responsible for strong binding (Koch et al., 1991).

Supervised machine learning methods have been applied in order to recognise interactions based solely on primary structure (Bock and Gough, 2001; Ofran and Rost, 2003). Using a combination of different machine learning techniques while focusing on sequence neighbours of a target residue, Yan et al. (2004) were able to identify interface residues on the basis of sequence information with an averaged accuracy of 72%.

Since, as previously claimed, the selective pressure on functional surface regions are known to be quite high, amino acids that contribute predominantly to the binding force should be highly conserved if the interacting function of an interface region is to be maintained during evolution. Comparing protein sequences among different species gives hint to so called evolutionary traces which can be used for interface prediction (Lichtarge et al., 1996; Lichtarge and Sowa, 2002).

The co-evolution of interacting proteins can be tracked closely by quantifying the degree of co-variation between pairs of residues from these proteins (correlated mutations). These positions may correspond to compensatory mutations that stabilise the mutations in one protein with changes in the other in order to further ensure the interaction function. Those correlated mutations can be detected by species-spanning sequence comparisons and used for the prediction of binding partners and the amino acids involved in interactions (Pazos et al., 1997; Valencia and Pazos, 2003; Bradford

and Westhead, 2003). The main limitation of this sometimes called "in silico two-hybrid" approach is the need for complete alignments with a good coverage of species common to the two proteins under study. This limitation arises as a direct consequence of the hypothesis of co-evolution, which naturally requires the simultaneous study of the corresponding protein pairs in each genome.

1.3.2.3 Structure based methods

Different varieties of approaching the problem of theoretical protein-protein interaction prediction depend on the amount of data available as input and the final aim of the prediction. The common feature of these methods is that they require the knowledge of structural data in order to calculate position specific geometrical, physical or chemical properties of the proteins in question.

- Prediction of interaction sites/interface regions

  These methods require the structural data of a single protein as input and will in return predict those residues or areas of the surface which are most likely part of an interface to other proteins.

  The bioinformatics tool ISPRED (Fariselli et al., 2002) uses evolutionary conservation along with surface disposition as descriptors to train a neural network (NN) based system. The NN is finally able to detect in average 73% of the residues involved in protein-protein interactions correctly within a selected database of heterodimers.

  The protein interaction prediction programs PROMATE (Neuvirth et al., 2004) and PPI-PRED (Bradford and Westhead, 2005) follow a different approach, aiming for the prediction of contiguous interface regions rather than interface residues.

  PROMATE has been developed using an extensive optimisation procedure in order to create a contiguous scoring function from individual scoring schemes created for each surface patch examined. The individual scoring schemes in use are based on amino acid propensities, pairwise amino acid distribution, evolutionary conservation, secondary structure information, sequence distance, distribution of temperature factors, the number of water molecules in the crystal structure

and hydrophobicity. The final predictor is able to correctly predict 70% of the interfaces for a dataset of transient dimeric complexes.

PPI-PRED uses evolutionary conservation and surface disposition (respectively solvent accessibility) along with further criteria such as the interface propensity of individual residue types, electrostatics, hydrophobicity and surface topography to train a Support Vector Machine (SVM) on this classification problem. For 76% of the interfaces of a selected dataset of homo- and heterodimeric complexes, a surface patch could be correctly predicted, showing at least 20% of correctly predicted interface residues while covering a minimum of 50% of the interface.

- Prediction of the complex structure

  Since the experimental determination of the complex structure (cf. 1.3.1.2 on page 14) is often disproportionate in difficulty to the determination of the protomers, the computational prediction of the 3D structure of a protein-protein complex from structural information of the protomers is of great interest. The process of the computational prediction of the complex structure, respectively to the prediction of the orientation of the complex subunits relative to each other in 3D space, starting from the structures of the protomers is called *protein-protein docking*. The search for candidate solutions in a docking problem is addressed in two essentially different approaches:

  (1) full solution space search

  This approach scans the entire solution space in a predefined systematic manner. Since an exhaustive search in the six dimensional conformational space (three degrees of freedom for rotation and translation each) using fully detailed information would be computationally too expensive, all existing approaches rely on reduced representations of the individual proteins.

  (2) gradual guided progression through solution space

  Only a part of the solution space is scanned in a partially random and partially criteria-guided manner. This approach consists mainly of simulations using Monte Carlo (MC), simulated annealing, molecular dynamics (MD), as well as evolutionary algorithms such as genetic algorithms (GA) and Tabu

search. Again, the simulation of many particle systems such as a protein-protein complex in solution over a sufficient amount of time is limited by currently available computation power.

Despite the high computational cost for these two general approaches above, all those methods rely on scoring or fitness functions to either evaluate the generated conformations (1) or guide the search (2).

Protein-protein docking using a full solution space search is particularly important for this work and is thus decribed in greater detail in the next section (section 1.4).

1.4 Protein-protein docking

The term protein-protein docking refers to the computational prediction of how two proteins interact; more precisely to the prediction of the orientation of the complex subunits relative to each other in 3D space. The fundamental basis from which all docking approaches emerged and still vastly rely on is the assumption of complementarity. Besides the question which properties complement each other, the usability of such complementaries in a docking procedure heavily depends on their nature. Of particular importane is the question whether the respective complementarity is implicitly present before the formation of the complex structure or whether it is induced by or during the association of the complex partners (see also subsection 1.4.2.3 on page 26).

1.4.1 The rigid body approach

Although there is no doubt that proteins are dynamical biological macromolecules, a large number of docking procedures published so far treat the individual proteins as rigid bodies in what is known as the *rigid body approach* or *rigid body docking*. Docking is computationally difficult because there are various ways of assembling two molecules (three translational and three rotational degrees of freedom). The number of possibilities grows exponentially with the size of the components. This is because a similar exponential growth is given for every additional degree of freedom that is introduced into the molecule in order to allow for internal movements, thus representing

protein flexibility. The combinatorial problem increases rapidly to such an amount that implementing full conformational flexibility into a search stage of a docking process is infeasible. The computational problem is even more profound when considering protein flexibility and the increasing demand to screen large databases.
There have been various approaches to incorporate protein flexibility into docking procedures which will be discussed in section 1.4.3 on page 31.

1.4.2 Principle steps of a docking procedure

Each docking method can be divided into four major steps (Halperin et al., 2002) consisting of

(i) representation of the system,

(ii) conformational space search,

(iii) scoring and ranking of potential solutions and

(iv) refinement of accepted solutions

which will be individually addressed to in the following subsections.

1.4.2.1 Representation of the system

Since interactions between proteins are mainly transmitted by those amino acids lying on the surface of the complex partners, any representation for the docking problem likewise focuses on descriptions of the protein surface. The basic description of the protein surface is given by the atomic representation of exposed residues. Such a representation in "atomic detail" is generally avoided because most algorithms scale with the number of representative points in three dimensional space and therefore, mathematical models of surface representation have been developed which offer a sparse distribution of surface points while simultaneously storing as much information as possible.
One frequently used approach originated from the pioneering work of Katchalski-Katzir et al. (1992) and Jiang and Kim (1991), where the proteins in question are mapped on a three dimensional grid of defined spacing with the spacing of the grid

determining the level of detail of the resulting lattice representation of the protein. Another popular approach is the surface represented by its geometric features for which Connolly (Connolly, 1983) laid the foundation with the developed method of protein surface analysis that bears his name since. The Connolly surface consists of that part of the van der Waals surface of the atoms that is accessible to the probe sphere (contact surface) connected by a network of convex, concave, and saddle shape surfaces that smooths over the crevices and pits between the atoms. Based on the Connolly analysis, the surface may be described by sparse critical points(Lin et al., 1994), defined as the projection of the gravity center of a Connolly face.

Parallel slices of the Connolly analysis can be transformed into a polygon to be used in a rigid surface matching (Ausiello et al., 1997). Jiang and Kim (1991) combine two representations of the molecule: surface dots with attached surface normals as proposed by Connolly, and volume (interior) and surface cubes, the latter containing two to three surface dots each.

Furthermore, volumetric and surface-based techniques for computing shape properties of molecular surfaces can be used. Several scalar and vector surface properties are gained, such as the Gaussian and mean curvature, principal curvatures, and principal curvature directions (Duncan and Olson, 1993). An extension to these methods is given by the description of protein surfaces using spherical harmonic functions where each protein surface shape is represented by a "double skin" model that describes thin regions of space exterior and interior to the molecular surface. Each skin is represented as a Fourier series expansion of real orthogonal radial and spherical harmonic basis functions (Ritchie and Kemp, 2000).

1.4.2.2 Conformational space search

Once the proteins, or rather their surfaces have been transformed into a mathematical surface representation, all possible orientations of the two individual subunits to each other have to be generated. The 3D structures of protein complexes reveal a close geometric and chemical match between those parts of the molecular surfaces that are in contact. Hence, the shape and other physical characteristics of the surfaces largely determine the nature of the specific interaction. Furthermore, in many cases the 3D structures of the components of the complex closely resemble those of the molecules in

their uncomplexed state (Lo Conte et al., 1999). Geometric matching is therefore likely to play an important part in determining the structure of the complex. All docking algorithms therefore search the conformational space for those structures revealing high correlations respectively complementaries of the adjacent surface areas.

- FFT-docking

  Fast Fourier Transformations (FFT) were first applied to the docking problem by Katchalski-Katzir et al. (1992) who also introduced the grid representations for proteins together with Jiang and Kim (1991) (cf. section 1.4.2.1 on page 23). The conformational space is searched for conformations in which the 3D grids representing the proteins are overlapping. Numerical values are assigned to the individual grid cells in order to control the desired overlap (cf. section 2.2.1 on page 41). The transformation of the proteins, respectively the established surface representations, into Fourier space reduces the dimensionality for the conformational space search from 6N to 3N. All possible conformations can be calculated in three dimensions simultaneously in Fourier space, thus reducing the computational cost effectively and making the effort of transforming each proposed conformation in the first 3N coordinates into Fourier space and back worthwhile. FFT docking algorithms are used in a large variety of docking programs nowadays (e.g. CKORDO (Zimmermann, 2002), MOLFIT (Ben-Zeev et al., 2005), DOT (Mandell et al., 2001), ZDOCK (Chen et al., 2003a), BDOCK (Huang and Schroeder, 2005), GRAMM (Vakser et al., 1999), FTDOCK (Gabb et al., 1997)). They rely on a grid representation of the protein subunits and can be considered as extensions of the initial approach by Katchalski-Katzir et al. (1992).

- Geometric hashing

  Conformational space search via geometric hashing is the transfer of a technique originally developed for object recognition problems in computer vision (Norel et al., 1994), in which the geometric hashing paradigm is adapted to a central problem in molecular biology. Using an indexing approach based on a transformation invariant representation, the algorithm efficiently scans groups of surface dots (or atoms) and detects optimally matched surfaces. Main advantage

for such a method is the ability to pre-calculate and store the transformation invariant representation - basically a transformation of the entire system into internal coordinates for every relevant set and combination of regarded surface points - such that the actual conformational space search is comparably fast.

1.4.2.3 Scoring and ranking of potential solutions

The search of the conformational space, be it complete or partially guided, typically yields a vast number of proposed complex conformations. From these proposed conformations ideally those have to be selected that show the highest similarity to the native complex conformation. This is the task of the so called scoring or ranking step of protein-protein docking. During this step, a numerical value is assigned to each of the proposed conformations according to a mathematical scheme and the individual cases are thereafter resorted according to their assigned numerical values. Theoretically, free-energy simulation can be a reliable discrimination to check the solutions. However, it is not practical to use such an approach in docking searches (Pearlman and Charifson, 2001) due to the vast computational effort of such calculations. Instead, ranking schemes are mostly used to distinguish between near-native solutions and others within a reasonable computation time. Two types of ranking schemes can generally be distinguished according to their sorting order. If high numerical values represent the desired outcome, the ranking scheme is classified as a scoring function, while energy or cost functions use low numerical values (often of negative sign) to represent the desired outcome. In the following, a list of possible criteria will be given that can be used to establish ranking schemes for protein-protein docking.

- Geometric complementarity/correlation

  As stated previously (cf. section 1.4 on page 22), docking methods vastly rely on the assumption of complementarity. Usually, geometric shape complementarity constitutes the first and most important scoring scheme and is generally the one which is directly computed while performing the scan of the conformational space. The main reason for this is the assumption that large parts of the energy gained upon complex formation result from the hydrophobic effect (Honig and Nicholls, 1995). Since hydrophobic forces are of short ranged nature, the complex partners should be in short distance to each other. Furthermore, as few gaps as

possible should be present in the interface area, which in consequence leads to the precondition that the two complex partners should exhibit corresponding radii of curvature on macroscopic and microscopic scale on their surfaces. Only then the geometric fit - the confrontation of concavities on one side of the interface with convexities on the other - will be established. Such predefined maps of the complex partner on the surface of a protomer can indeed be identified on a series of protein-protein complexes already in the unbound state (Betts and Sternberg, 1999). For some selected problems in docking, geometric shape complementarity is already sufficient in order to establish a scoring function that sorts and yields near native conformations in the top rank(s) (see subsection 1.4.4 on page 33). However, these cases clearly are the exception and so further ranking/scoring criteria have to be considered.

- Physico-chemical complementarities/correlations

  - Electrostatics

    All electric charges in a protein contribute to the characteristic field of charge on the protein surface. During the transition from the unbound to the complexed state, the interface area moves from an (in vivo usually aqueous) environment with a rather high dielectric constant to an environment with a much lower dielectric constant which resembles more the protein core than the surface. This requires that the geometric fit of the contact area is tight enough to exclude solvent molecules. This drastic change in the dielectric constant leads to an increased loss in energy for every charge that is not compensated by its counterpart and explains the need for charge complementarity across interfaces (Gabb et al., 1997; Sheinerman and Honig, 2002). Studies have shown, that charge complementarity is nevertheless insignificantly small in a number of protein-protein interfaces and that rather the electrostatic correlation of the surface electrostatic potential is of significance (McCoy et al., 1997). The easiest approach for the calculation of electrostatic potentials in protein-protein interactions takes only the sum of the potentials of individual point charges into account, using force fields primarily based on the classical Coulomb potential and extensions to the latter. A more correct but also computationally much

more expensive way is to calculate the complete field of charge of the proteins using continuum electrostatic models e.g. via Poisson-Boltzmann approaches (Jackson and Sternberg, 1995; Gabdoulline and Wade, 1998; Mandell et al., 2001; Neves-Petersen and Petersen, 2003).

  - Hydrophobicity
    Formation of hydrophobic contacts across a newly formed interface is energetically favourable, especially when the drastic change in the dielectric constant in the interface area upon complex formation (see above) is taken into account (Scarsi et al., 1999). The extent of such hydrophobic complementarity depends on the size of the interface. Thus, the non-polar portions of large interfaces are more often juxtaposed to each other than non-polar portions of small interfaces (Berchanski et al., 2004).

  - Desolvation
    The desolvation free energies required to transfer atoms from the surface of a protein to a protein's interior (e.g. the interface of a protein-protein complex) are particularly hard to calculate analytically, since large parts of the desolvation free energy will actually be contributed by entropic terms. Appropriate estimations and empirical measures for desolvation free energies are such used as scoring functions (Miyazawa and Jernigan, 1996; Zhang et al., 1997; Wang and Wade, 2003).

  - Hydrogen bonds
    Especially charged interfaces are known to show certain hydrogen bonding patterns. Based on the assumption that these hydrogen bonding patterns of complementary hydrogen bond donors and acceptors on the surface or the protomers are predefined already in the uncomplexed state, this complementarity can be used in order to score proposed docking solutions (Meyer et al., 1996; Krämer, 2001; Fernández and Scheraga, 2003).

- empirical scoring schemes

  - Knowledge based scoring functions
    Especially electrostatic as well as hydrophobic complementarities can be implicitly expressed by the distribution patterns of aminoacids or atoms

on a proteins surface. By assuming a Boltzmann relation between the frequency of occurrence and the energy of a certain state of the molecule as well as an additive relation for individual contributions to the overall energy, the binding free energy of a complex can be estimated (Sippl, 1990). The resulting pseudo-energies can be used to score docking solutions. Thus a large variety of these empiric interaction potentials, sometimes also described as probability density functions (PDF), have been developed via calculation of frequencies of occurrences of interaction pairs (be that amino acids, atom groups or atoms) (Miyazawa and Jernigan, 1996; Melo and Feytmans, 1997; Moont et al., 1999; Verdonk et al., 2001; Grimm, 2003; Zhang et al., 2004).

– "Biological" scoring functions
Besides the derivation of probability density functions from biological observations there exist various other methods that allow the usage of such knowledge for the scoring of docking results. In principle, all the methods described previously for the computational prediction of protein-protein interaction can be utilised as scoring functions for docking (see section 1.3.2 on page 17). Examples are:

  * the use of evolutionary information by the adoption of sequence conservations to re-rank docking solutions (Halperin et al., 2004; Duan et al., 2005; Heuser et al., 2005; Tress et al., 2005; Aytuna et al., 2005),
  * the utilisation of sequence to structure relations to identify homologous domains known to interact (Heuser et al., 2005) or establish protein family specific residue interface propensities (Huang and Schroeder, 2005),
  * considering the buried surface area and the gap volume in order to estimate the tightness of binding (Gardiner et al., 2003; Gottschalk et al., 2004; Huang and Schroeder, 2005).

Besides the application of scoring schemes to docking solutions, the incorporation of external knowledge into docking becomes more and more important. Such external knowledge, e.g. the knowledge derived from H/D labeling mass spectrometry experiments that a certain residue in one of the protomers has to be part of the interface

(cf. 1.3.1.1 on page 12), can be efficiently used not only to score docking solutions but also to guide the conformational space search efficiently (Ben-Zeev and Eisenstein, 2003). If external knowledge provided can tell which areas of the proteins surface are definitely participating in the interaction or which areas can be generally ruled out, the complete docking process as well as the subsequent re-ranking can be shortened and eased drastically.

1.4.2.4 Refinement of accepted solutions

Predictions generated by a rigid body docking algorithm can only be as good as the underlying assumption. Proteins are no rigid bodies and thus are likely to undergo conformational changes when transferred from one environment (protomer in solvent) to another (complex). The range of such changes reaches from mere side chain rearrangement via movement of flexible loop regions to shear and hinge bending between domains (Betts and Sternberg, 1999; Smith et al., 2005b). These conformational changes upon complex formation cannot be captured effectively by most docking algorithms. A divide-and-conquer strategy is widely accepted in the field of docking, with initial-stage algorithms focused on retaining near-native structures (also called hits) (cf. 1.4.2.2 on page 24) in a reasonably short list of predictions and scoring functions (cf. 1.4.2.3 on page 26) aimed at ranking a hit at the top of the list. The actual task of a refinement algorithm for rigid-body docking is to allow for a finer re-ranking of those near-native structures that ranged on the top of the list in the previous scoring step.

If the native complex and the individual subunits submitted to a docking procedure are structurally not identical (see section 1.4.4 on page 33), the rigid body approach poses a severe limit on how close a near native docking solution can actually be brought to the ideal solution, respectively the native complex. Thus modern refinement algorithms focus on the simulation and approximation of possible conformational changes that occur upon complex formation. Since the refinement step is consequently only applied to a very limited number of docking solutions, more time consuming computational methods can be used. Among these methods are short position restrained molecular dynamics simulations (Gillilan and Lilien, 2004; Grünberg et al., 2004; Smith et al., 2005a,b), energy minimisation procedures (Jackson et al., 1998; Li et al., 2003b; Wiehe

et al., 2005; Carter et al., 2005) and the use of rotamer libraries (Jackson et al., 1998; Koch et al., 2002; Althaus et al., 2002; Carter et al., 2005) that explicitly account for possible conformational changes of side chains. Afterward, highly specific scoring functions are applied to evaluate the interaction energy. These steps are then repeated until convergence of the resulting (pseudo-)energy.

1.4.3 Incorporation of flexibility into protein-protein docking

Protein-protein association is often accompanied by changes in receptor and ligand structure. This interplay between protein flexibility and protein-protein recognition is currently the largest obstacle both to the understanding and to the reliable prediction of protein complexes. Besides of the incorporation of flexibility treatment in the final refinement step of a docking procedure (as described in the previous section), it is most sensible to include flexibility in the critical step of docking, the conformational space search. Only thus, protein assemblies which do not exhibit a predefined geometric fit and rather follow a transfer of the induced fit model to protein complexes than the "key and lock" hypothesis as originally proposed by Emil Fischer (Fischer, 1894) for enzyme substrate binding, can be correctly predicted.
Various approaches exist in order to integrate flexibility into docking algorithms while searching the conformational space based on the fact that flexibility can be adressed at several levels.

- Implicit treatment of flexibility in docking
  On an implicit level, flexibility can be treated by smoothing the protein surfaces or allowing some degree of interpenetration (*soft docking*) or by performing multiple docking runs from various conformations (*cross- or ensemble docking*)(Bonvin, 2006).

  - Soft docking
    Within the framework of the rigid body treatment, side chain flexibility is typically handled only implicitly by surface variability, with a soft belt of allowed (though sometimes penalised) intermolecular surface atom penetration. There also exist approaches to evade the problem of side chain reorientation by submitting rather coarse and simplified protein models to

the search step that do not contain side chain atoms at all (Vakser, 1996, 1995) or provide only partial information about the side chains, e.g. by cutting them off according to certain rules (Li et al., 2003a; Schneidman-Duhovny et al., 2005a) or replacing them by a limited number of pseudo atoms (Zacharias, 2003). However, these methods are rather auxiliary constructions for the problem of flexibility in protein-protein docking.

- Cross- or Ensemble docking
  Implementing full conformational flexibility into a search stage, separately docking a large number of conformers, is infeasible. A reasonable approach is to take account of ensembles of populations, generated prior to the docking, and dock the ensemble rather than single conformers. Depending on the strategy, docking an ensemble highlights the more conserved regions by, for example, assigning these larger weights, whereas lower weights may be given to regions of space visited more rarely. Experimentally, ensembles can be assembled by collecting all crystal structures binding to a certain ligand, or using NMR conformers (Halperin et al., 2002). Unfortunately, data fundamentals are quite low for protein-protein docking such that this approach is only feasible for a very limited number of examples. Nevertheless, it is possible to generate an ensemble of hypothetical conformations of the individual complex subunits prior to the docking. The creation of theoretical ensembles can be achieved via genetic algorithms (Taylor and Burnett, 2000), Monte Carlo algorithms (Gray et al., 2003), molecular dynamics methods (Smith et al., 2005a), multi-conformational superposition (Ma et al., 2005) or the detection of hinge regions (Schneidman-Duhovny et al., 2003, 2005a).

- Explicit treatment of flexibility in docking
  The inclusion of flexibility in docking is only possible when molecules are explicitly represented rather than via a mathematical simplified model (e.g. a grid). Since most of the currently used docking methods do not use a full representation of the molecule during the search of the conformational space, explicit treatment of flexibility in docking is typically handled during the refinement step(s) (see 1.4.2.4 on page 30). Generally, one can distinguish between the

incorporation of side chain and backbone flexibility.

- Amino acid side chain flexibility
  The general methods available are energy minimisations, often coupled with the use of position restrainied simulation methods and rotamer libraries. A few examples of methods along with the corresponding docking software names are listed below.
  - Monte Carlo optimization of sidechains (ICM-DISCO, (Fernández-Recio et al., 2003))
  - Molecular Dynamics simulated annealing (HADDOCK, (Dominguez et al., 2003))
  - Energy minimisation and multiple sidechain conformations using rotamer libraries (ATTRACT, (Zacharias, 2005))
  - Monte Carlo search that includes rigid-body displacements using rotamer libraries (ROSETTADOCK, Gray et al. (2003))
- Backbone flexibility
  Dealing with backbone flexibility in protein-protein docking is still an open challenge. The incorporation of explicit backbone conformational changes currently relies on molecular dynamics simulation techniques. A few examples of methods along with the corresponding docking software names are listed below.
  - Molecular Dynamics simulated annealing (HADDOCK, (Dominguez et al., 2003))
  - Guided docking which allows for some degree of backbone rearrangement ((Fitzjohn and Bates, 2003))

1.4.4 Docking problems and challenges

There are two different general case studies of protein-protein docking at different levels of complexity, for which the terminology of *bound docking* and *unbound docking* are commonly used.
Bound docking denotes the attempt of the computational reassembly of the subunits of a complex of known structure, often a cocrystallised complex structure, which have

previously been taken apart. The *bound docking* problem is generally regarded as solved, since most rigid body docking methods are able to find the native complex structure or an appropriate near-native solution with high accuracy (Vajda and Camacho, 2004). This is mostly due to the well established geometric fit between the binding partners.

Unbound docking, sometimes also called predictive docking, refers to the prediction of the native complex state from the unbound subunits. The structures of the unbound subunits have to be solved in a solvent accessible state (at least in the respective interface region; for a more detailed definition of an unbound docking case, see section 2.1 on page 36). Predictive docking is far more complex than bound docking. The additional complexity derives from conformational changes that take place between the bound and unbound structures.

In order to specify the nature of the docking problem in more detail, it is common for the usually binary dockings of receptor versus ligand to use a composed terminology. It denotes the binding states in which both the subunit structures are situated in the moment of their structures' solution, mostly in the order of receptor (usually the larger of the two protomers) followed by the ligand state. This leads to four different notations, listed here in the order of complexity of the problem: bound-bound, bound-unbound, unbound-bound and unbound-unbound docking.

The bound-unbound versions of docking result if only one of the subunits of the complex is actually available as individually crystallised structure. This often is the case since data fundamentals for docking are quite low and represent a problem presumably easier than unbound- but more difficult than bound docking. These cases are also known as *crossbound-docking*.

While bound docking is only of academic use that will allow for a fundamental answer to the question whether protein-protein interaction prediction is possible using a certain algorithm, unbound docking is much closer to a real world application. Since existing approaches to the unbound docking problem are quite diverse while the number of known test cases is relatively small, there is the risk of those methods being geared to the limited data fundamentals used in their design. In order to asses the quality of existing docking methods and provide an overview of the status quo of current research and performance in protein-protein interaction prediction, a comparative academic challenge has been brought to life. Role model for this docking challenge was the CASP

(Critical Assessment of protein Structure Prediction) challenge (Moult, 2005) which now exists for about a decade and focuses on the evaluation of predictions of protein structures from sequence information. In 2001, the CAPRI (Critical Assessment of PRedicted Interactions) challenge (Janin et al., 2003) was established offering an assessment of blind docking predictions to the research community. The results of the CAPRI challenge are published and summarised in a special issue of the journal PROTEINS (Méndez et al., 2003, 2005; Janin, 2005) every two years.

1.5 Aim of work

This work deals with the ranking or scoring problem of a protein-protein docking procedure. A protein-protein docking algorithm typically yields a vast number of potential solutions during the conformational space search. It is the aim of this work to establish new scoring functions for protein-protein docking as well as to find a way to sensibly combine these functions such that near-native solutions can be accurately detected and selected. The scoring scheme(s) should specifically be applicable to challenging unbound-unbound docking problems, where the geometric fit and its primary correlation functions are unsufficient or fail in the ranking of prospective candidates. Primary goal is to reduce the number of candidates for any further refinement steps. The method should be applicable to any underlying method of conformational space search while ensuring easy extensibility for future incorporation of further scoring schemes. The protein-protein docking calculations were executed with the docking software CKORDO developed in the workgroup. This work will focus on an extension and improvement of the software for a future postfiltering step.

2 Methods

"Though this be madness, yet there is method in it." [Hamlet]

William Shakespeare, 1564-1616.

2.1 Data fundamentals

From the currently more than 35,000 protein structures deposited in the Protein Data Bank (PDB) (March 2006), only very few (<1%) fulfil the criteria necessary for an unbound-unbound protein-protein docking test case. The relevant criteria are:

1. All subunits of the complex should also be found as individually crystallised structures in the PDB, with at least the required interface region in a solvent accessible state,

2. the co-crystallised complex should be a heteromultimeric complex,

3. the protein structures should not be hypothetical or modelled,

4. the resolution of the interface area should be complete and qualitatively as high as possible.

The search for suitable unbound docking test cases for a docking algorithm in the PDB is difficult since the PDB is a collection of flat files, each containing information about a single structure, with an insufficient number of attributes and without any relation between the files. A manual collection of known unbound-unbound docking examples from the literature was therefore performed.
Table 2.1 on the facing page gives an overview of the collected examples along with the reference in which this docking test case has previously been used. For each of the binary docking test cases, the PDB identifier for the complex and the unbound units are given. Furthermore, the chains involved are specified along with the number of residues in the respective chain(s).

The resulting collection of test cases was checked for redundancy. This was done by deriving the sequences from the structures and performing a full factorial BLAST (Altschul et al., 1990) search. Only those unbound-unbound docking test cases were retained that showed a maximum of 75% sequence positives to any other complex in the dataset while exhibiting a minimum of 75% positives between the complex and the respective unbound units. This resulted in a total of 33 test cases which classify into 21 Enzyme-Inhibitor complexes, four Antibody-Antigen complexes, four "other" complexes (not belonging to either of the previous groups) and four "difficult" test cases. The test cases classified as difficult are those that undergo drastic conformational changes upon complex formation and thus represent the biggest challenge to a rigid-body docking algorithm. This collection of unbound-unbound docking test cases represents a on sequence level non-redundant version of the first ever published benchmark for protein-protein docking (Chen et al., 2003b), using the same classification scheme.

Table 2.1: Unbound-unbound protein-protein docking examples as collected from the literature.

co-crystallised complex		unbound 1 (receptor)			unbound 2 (ligand)			
PDB-ID	chain(s)	PDB-ID	Chain(s)	#Res	PDB-ID	Chain(s)	#Res	Ref.
Enzyme-Inhibitor/Enzyme-Substrate complexes (21)								
1ACB	E:I	5CHA	A	237	1CSE	I	63	i
1AVW	A:B	2PTN	-	223	1BA7	A	165	i
1BRC	E:I	1BRA	-	223	1AAP	A	56	i,j
1BRS	A:D	1A2P	B	108	1A19	A	89	a,c,e,i,j
1BVN	P:T	1PIF	-	495	2AIT	-	74	j
1CGI	E:I	1CHG	-	230	1HPT	-	56	b,c,d,e,i,j
1CHO	E:I	5CHA	A	237	2OVO	-	56	a,c,e,g,i,j
1CSE	E:I	1SCD	-	274	1ACB	I	63	i
1DFJ	I:E	2BNH	-	456	7RSA	-	124	i,j
1FSS	A:B	2ACE	-	527	1FSC	-	61	a,c,e,f,i,j
1MAH	A:F	1MAA	B	536	1FSC	-	61	c,h,i
1PPF	E:I	1PPG	E	218	2OVO	-	56	a
1TGS	Z:I	2PTN	-	223	1HPT	-	56	i
1UGH	E:I	1AKZ	-	223	1UGI	A	83	i,j
2KAI	AB:I	2PKA	XY	232	6PTI	-	57	c,d,i,j
2PTC	E:I	2PTN	-	223	6PTI	-	57	a,c,d,i,j
2SIC	E:I	1SUP	-	275	3SSI	-	108	c,d,i,j
2SNI	E:I	1SUP	-	275	2CI2	I	65	c,d,e,i,j
2MTA	LH:A	2BBK	LH	480	1AAN	-	105	i

...continued on next page

Table 2.1 – continued from previous page

co-crystallised complex		unbound 1 (receptor)			unbound 2 (ligand)			
PDB-ID	chain(s)	PDB-ID	Chain(s)	#Res	PDB-ID	Chain(s)	#Res	Ref.
2PCB	A:B	1CCP	-	293	1HRC	-	105	[e,j]
2PCC	A:B	1CCA	-	291	1YCC	-	107	[i,j]
Antibody-antigen complexes (4)								
1AHW	DE:F	1FGN	LH	428	1BOY	-	211	[i,j]
1DQJ	AB:C	1DQQ	AB	424	3LZT	-	129	[i]
1VFB	AB:C	1VFA	AB	224	1LZA	-	129	[b,c,f,j]
1WEJ	LH:F	1QBL	LH	433	1HRC	-	105	[i,j]
'Other' complexes (4)								
1AVZ	B:C	1AVV	-	99	1SHF	A	59	[i]
1L0Y	A:B	1BEC	-	238	1B1Z	A	218	[i]
1WQ1	G:R	1WER	-	324	5P21	-	166	[i]
1BDJ	A:B	3CHY	-	128	2A0B	-	118	[i]
'Difficult test cases' (4)								
1BTH	LH:P	2HNT	LCEF	292	6PTI	-	57	[i]
1FIN	A:B	1HCL	-	294	1VIN	-	252	[i]
1FQ1	B:A	1B39	A	290	1FPZ	F	178	[i]
1GOT	BG:A	1TBG	AE	408	1TAG	-	314	[i]

These examples were collected from a total of ten different literature resources [a−j]. During the course of this work, a new, much larger protein-protein docking benchmark was published (Mintseris et al., 2005). For this benchmark, the PDB has been parsed for putative docking test cases using new quality and redundancy criteria. This benchmark now holds a total of 84 non-redundant docking test cases. These test cases consist of transient native complexes, which are structurally non-redundant, along with those unbound structures that have the highest possible sequence identity to the bound interactors, while consisting of those crystal structures with the lowest possible resolution and the fewest residues with missing electron density. Structural redundancy was avoided by using the Structural Classification Of Proteins SCOP (Andreeva et al., 2004) hierarchical domain classifications, taking family-family pairs

[a] Camacho and Vajda (2001)
[b] Gardiner et al. (2001)
[c] Chen and Weng (2002)
[d] Gabb et al. (1997)
[e] Palma et al. (2000)
[f] Heifetz et al. (2002)
[g] Lorber et al. (2002)
[h] Mandell et al. (2001)
[i] Halperin et al. (2002)
[j] Chen et al. (2003b)

as non-redundant unit. For the obtained 84 complexes, Mintseris et al. performed an FFT-docking in order to classify them according to their expected difficulty for most docking methods. The number of high-quality hits, defined by interface root mean square deviation (RMSD) and the fractions of native and non-native contacts as used in the CAPRI challenge (Méndez et al., 2003), were employed for the classification. 63 test cases have been classified as "rigid-body" docking problems, 13 are listed as "medium-difficulty" while eight "difficult" examples are given. Besides this classification, the docking test cases are grouped into Enzyme-Inhibitor/Enzyme-Substrate complexes, Antibody-Antigen complexes and "Other" complexes for those not belonging to either of the two previous groups.

Table 2.2: Protein-protein docking benchmark 2.0 (Mintseris et al., 2005).

co-crystallised complex		**unbound 1 (receptor)**			**unbound 2 (ligand)**		
PDB-ID	chain(s)	PDB-ID	Chain(s)	#Res	PDB-ID	Chain(s)	#Res
Rigid-body (63)							
Enzyme-inhibitor / Enzyme-substrate complexes (21)							
1AVX	A:B	1QQU	A	223	1BA7	B	169
1AY7	A:B	1RGH	B	96	1A19	B	89
1BVN	P:T	1PIG	-	495	1HOE	-	74
1CGI	E:I	2CGA	B	245	1HPT	-	56
1D6R	A:I	2TGT	-	223	1K9B	A	58
1DFJ	I:E	2BNH	-	456	9RSA	B	124
1E6E	A:B	1E1N	A	455	1CJE	D	107
1EAW	A:B	1EAX	A	241	9PTI	-	58
1EWY	A:C	1GJR	A	295	1CZP	A	98
1EZU	AB:C	1ECZ	AB	284	1TRM	A	223
1F34	A:B	4PEP	-	326	1F32	A	127
1HIA	AB:I	2PKA	XY	232	1BX8	-	49
1MAH	A:F	1J06	B	533	1FSC	-	61
1PPE	E:I	1BTP	-	223	1LU0	A	29
1TMQ	A:B	1JAE	-	470	1B1U	A	117
1UDI	E:I	1UDH	-	228	2UGI	B	83
2MTA	HL:A	2BBK	JM	480	2RAC	A	105
2PCC	A:B	1CCP	-	293	1YCC	-	107
2SIC	E:I	1SUP	-	275	3SSI	-	108
2SNI	E:I	1UBN	A	274	2CI2	I	65
7CEI	B:A	1M08	B	131	1UNK	D	87
Antibody-antigen complexes (9)							
1AHW	AB:C	1FGN	LH	428	1TFH	A	202

...continued on next page

Table 2.2 – continued from previous page

co-crystallised complex		unbound 1 (receptor)			unbound 2 (ligand)		
PDB-ID	chain(s)	PDB-ID	Chain(s)	#Res	PDB-ID	Chain(s)	#Res
1BVK	DE:F	1BVL	BA	224	3LZT	-	129
1DQJ	AB:C	1DQQ	CD	424	3LZT	-	129
1E6J	HL:P	1E6O	HL	429	1A43	-	72
1JPS	HL:T	1JPT	HL	425	1TFH	B	182
1MLC	AB:E	1MLB	AB	432	3LZT	-	129
1VFB	AB:C	1VFA	AB	224	8LYZ	-	129
1WEJ	HL:F	1QBL	HL	433	1HRC	-	105
2VIS	AB:C	1GIG	LH	431	2VIU	ACE	960
"Other" complexes (22)							
1A2K	AB:C	1OUN	AB	246	1QG4	A	202
1AK4	A:D	2CPL	-	164	1E6J	P	210
1AKJ	AB:DE	2CLR	DE	375	1CD8	AB	228
1B6C	B:A	1IAS	A	330	1D6O	A	107
1BUH	A:B	1HCL	-	294	1DKS	A	76
1E96	B:A	1HH8	A	192	1MH1	-	183
1F51	AB:E	1IXM	AB	343	1SRR	C	121
1FC2	D:C	1FC1	AB	414	1BDD	-	60
1FQJ	A:B	1TND	C	316	1FQI	A	133
1GCQ	C:B	1GCP	B	67	1GRI	B	211
1GHQ	A:B	1C3D	-	294	1LY2	A	130
1HE1	C:A	1MH1	-	183	1HE9	A	131
1I4D	AB:D	1I49	AB	402	1MH1	-	183
1KAC	A:B	1NOB	F	185	1F5W	B	121
1KLU	AB:D	1H15	AB	369	1STE	-	238
1KTZ	B:A	1M9Z	A	105	1TGK	-	112
1KXP	D:A	1KW2	B	453	1IJJ	B	371
1ML0	AB:D	1MKF	AB	742	1DOL	-	71
1QA9	A:B	1HNF	-	179	1CCZ	A	171
1RLB	ABCD:E	2PAB	ABCD	456	1HBP	-	175
1SBB	B:A	1SE4	-	239	1BEC	-	238
2BTF	A:P	1IJJ	B	371	1PNE	-	140
Antibody-antigen complexes (Crossbound) (11)							
1BJ1	HL:VW	1BJ1	HL	431	2VPF	GH	189
1FSK	BC:A	1FSK	BC	434	1BV1	-	159
1I9R	HL:ABC	1I9R	HL	434	1ALY	ABC	438
1IQD	AB:C	1IQD	AB	408	1D7P	M	159
1K4C	AB:C	1K4C	AB	431	1JVM	ABCD	394
1KXQ	A:H	1PPI	-	496	1KXQ	H	120
1NCA	HL:N	1NCA	HL	435	7NN9	-	388
1NSN	HL:S	1NSN	HL	427	1KDC	-	137
1QFW	HL:AB	1QFW	HL	224	1HRP	AB	196
1QFW	IM:AB	1QFW	IM	229	1HRP	AB	196
2JEL	HL:P	2JEL	HL	435	1POH	-	85
Medium-difficulty (12)							

...continued on next page

Table 2.2 – continued from previous page

co-crystallised complex		unbound 1 (receptor)			unbound 2 (ligand)		
PDB-ID	chain(s)	PDB-ID	Chain(s)	#Res	PDB-ID	Chain(s)	#Res
Enzyme-inhibitor / Enzyme-substrate complexes (2)							
1ACB	E:I	2CGA	B	245	1EGL	-	70
1KKL	ABC:H	1JB1	ABC	471	2HPR	-	87
Antibody-antigen complexes (Crossbound) (1)							
1BGX	T:HL	1CMW	A	817	1AY1	HL	423
"Other" complexes (9)							
1GP2	BG:A	1TBG	DH	405	1GIA	-	310
1GRN	B:A	1RGP	-	189	1A4R	A	190
1HE8	A:B	1E8Z	A	839	821P	-	166
1I2M	B:A	1A12	A	401	1QG4	A	202
1IB1	AB:E	1QJB	AB	460	1KUY	A	166
1IJK	BC:A	1FVU	AB	254	1AUQ	-	208
1K5D	AB:C	1RRP	AB	338	1YRG	B	343
1M10	B:A	1M0Z	B	266	1AUQ	-	208
1N2C	ABCD:EF	3MIN	ABCD	491	2NIP	AB	289
1WQ1	G:R	1WER	-	324	6Q21	D	171
Difficult (8)							
"Other" complexes (7)							
1ATN	A:D	1IJJ	B	371	3DNI	-	258
1DE4	CF:AB	1CX8	AB	1278	1A6Z	AB	371
1EER	BC:A	1ERN	AB	416	1BUY	A	166
1FAK	HL:T	1QFK	HL	348	1TFH	B	182
1FQ1	B:A	1B39	A	290	1FPZ	F	178
1IBR	B:A	1F59	A	440	1QG4	A	202
2HMI	AB:CD	1S6P	AB	979	2HMI	CD	434
Antibody-Antigen complexes (1)							
1H1V	A:G	1IJJ	B	371	1D0N	B	729

2.2 Docking algorithm

2.2.1 Ckordo

All the docking calculations in this work have been conducted using the Ckordo docking software as developed by Zimmermann (2002). The algorithm is an enhancement of the Kordo algorithm (Meyer et al., 1996) which itself is based on a method developed by Katchalski-Katzir et al. (1992) during which the correlation of two discretised protein surfaces is calculated in Fourier space with increased efficiency as compared to the calculation in direct space. This rigid-body docking algorithm is suitable only for binary docking problems. The two subunits to be docked are mapped on a three dimensional grid. The receptor, typically the larger of the subunits to be docked, is

position restrained in space, while the ligand, respectively the grid representation of the ligand, is rotated by discrete angle increments for a full search of the rotational space. CKORDO calculates the surface geometry correlation as well as a hydrophobic and an electrostatic potential for every orientation/conformation.
In order to calculate the geometric correlation scores, single integer values are assigned to every grid cell in the 3D grid representations. The position restrained protein A (receptor) is mapped on three different types of grid cells (equation (2.1)), with a defined protein interior, surface layer and those cells not explicitly occupied by any atom of the protein (free space in the grid beyond the proteins measurements). For moving protein B (ligand), only two cell types are distinguished: protein interior cells and cells "outside" of the protein (see equation (2.2)).

$$f_{A_{i,j,k}} = \begin{cases} 1 & \text{surface layer} \\ \rho & \text{protein interior} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

$$f_{B_{i,j,k}} = \begin{cases} 1 & \text{protein interior} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where:

f_A, f_B	:	numerical values assigned to regarded grid cell
i, j, k	:	internal coordinates of the 3D grid

The final geometric correlation is calculated by multiplication of overlapping grid cells of the two individual grids according to equation (2.3). In order to punish undesired

$$f_{C_{\alpha,\beta,\gamma}} = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{L} f_{A_{i,j,k}} \cdot f_{B_{i+\alpha,j+\beta,k+\gamma}} \tag{2.3}$$

where:

$f_{C_{\alpha,\beta,\gamma}}$	:	geometric correlation score for given orientation (rotation fixed, only translational dependencies included)
i, j, k	:	internal coordinates of the 3D grid
N, M, L	:	maximum dimensions of the grid in direction of i,j,k, respectively
α, β, γ	:	components of translation vector in internal units of grid

orientations which lead to a large overlap of the two proteins, especially an overlap of protein B with the interior of protein A, ρ is usually a negative integer value. In this work, the CKORDO default value of $\rho = -6$ is used. An overlap of protein B with the surface layer of protein A will lead to positive contributions, an overlap of protein B with the interior of protein A to negative contributions while orientations without overlap will lead to zero values. It is important to note here, that the correlation function as described above is uni-directional. The ligand is not surrounded by a surface layer in the grid representation, thus the problem is not symmetric. Interchanging receptor with ligand and vice versa will not lead to identical results.

The calculation of the geometric correlation score in Fourier space reduces the algorithmic complexity of the underlying problem from $O(N^6)$ to $O(N^3 log(N^3))$.

Electrostatic and hydrophobic correlation terms are calculated in a likewise manner. Herefore, pseudo coulomb potentials as well as hydrophobicity terms taken from the AMBER95 (Pearlman et al., 1995) force field are mapped to separate grid representations of the proteins and the correlation scores calculated in Fourier space.

For the maxima of geometric correlation the value for a pairwise atom-atom contact potential is calculated as well as optionally the buried surface area and the gap volume via the external programs DSSP (Kabsch and Sander, 1983) and SURFNET (Laskowski, 1995).

2.2.2 RMSD calculations

The main quality criterion for any docking calculation is the root mean square deviation (RMSD) as a measure for structural similarity calculated between the putative complex orientations yielded by the docking software and a reference state. For unbound docking this reference state can either be the native complex or the unbound units as fitted on the native complex. Since the primary sequence of the unbound units often differs from the corresponding native complex, a structural alignment algorithm is necessary to assign corresponding residues between unbound and native complex. The CKORDO algorithm does not facilitate a structural alignment, such that the RMSD calculations can only be conducted when using the unbound units as fitted on the native complex as a reference state. Consequently, all the RMSD calculations in this work are executed in the same manner. The calculation of the RMSD with respect to the fitted

unbound complex is reasonalble since the unbound units fitted on the native complex represent the best possible solution for a rigid body docking, where flexibility is not explicitly handled in the conformational space search. For identification of a near-native protein complex structure, structural similarity in the interface or contact region of the protomers is of far greater importance than structural similarity in regions which are not in contact with each other. Therefore, only the RMSD of interface atoms is used in this work, where all those atoms are defined as interface atoms for which an atom of the complex partner can be found within a threshold of 6Å euclidean distance. Since flexible side chains are likely to undergo conformational changes when transferred from one environment (unbound subunit as crystallised in solution) to another (interface of a protein-protein complex), only C-alpha atoms are taken into account.
The root mean square deviation of interface C-alpha atoms ($\mathrm{RMSD}_i C_\alpha$) calculated between a putative complex orientation and the unbound units as fitted on the native complex will be the major criterion to judge the quality of protein-protein docking results in this work.

2.3 The GRID software package

The term GRID specifies a software package which is widely used especially in pharma industry. It has been initially developed by Peter Goodford (Goodford, 1985) for the identification of non-covalent interaction forces between a molecule of known 3D-structure (*target*), usually a biological macromolecule, and a user defined chemical group (*probe*). Various energetic potential hyperplanes can such be generated and used for the identification of binding sites for the respective probe in the target structure.
The program GRID advanced to a standard tool for the use of macromolecular structure information, mostly of protein structures, in the development of new therapeutical agents. Drug molecules should be designed such that they exactly match the structure of a desired target molecule geometrically and also chemically. GRID offers the possibility to judge the energetic and geometric correlation of a protein-pharmacophore-system.
For this, the target molecule is wrapped in a three dimensional grid of defined spacing which expands beyond the maximum extensions of the target by a predefined measure. For every grid point that does not explicitly collide with the van der Waals radius of one

of the target atoms, an interaction energy is calculated for the user defined chemical probe (usually a relatively small molecular fragment, often just single atoms).
Each probe is characterised in its chemical and physical properties by the following parameters: The van der Waals radius r, the effective number of electrons N_{eff}, the polarisability α, the electrostatic charge Q, the minimum energy value E_{min} in the energy function E_{hb} that calculates the hydrogen bonding energy, the maximal number of donated hydrogen bonds JD, the maximal number of accepted hydrogen bonds JA as well as a numerical value $JTYPE$, the "hydrogen bonding type", which provides information about the preferred hydrogen bonding geometry (preferred angles between accepted and donated hydrogen bonds) of the probe. In a likewise manner, every single surface atom of the target is characterised by the parameters described. The interaction energy at a certain grid point with the coordinates x, y, z is calculated using the empirical energy functions (2.4 - 2.7):

$$E_{xyz} = \sum E_{lj} + \sum E_{el} + \sum E_{hb} \tag{2.4}$$

$$E_{lj} = \frac{A}{d^{12}} - \frac{B}{d^{6}} \tag{2.5}$$

$$E_{el} = \frac{pq}{K\zeta}\left[\frac{1}{d} + \frac{\frac{(\zeta-\epsilon)}{(\zeta+\epsilon)}}{\sqrt{d^2 + 4s_p s_q}}\right] \tag{2.6}$$

$$E_{hb} = \left[\frac{C}{d^6} - \frac{D}{d^4}\right]\cos^m\theta \tag{2.7}$$

Equation (2.5) represents the well known Lennard-Jones-potential. In this equation, d is the euclidean distance between two non covalently bound atoms for which the (Lenard-Jones) energy E_{lj} is described by parameters A and B. The values for A and B are calculated according to Hopfinger (1973) from the effective number of electrons N_{eff}, the polarisability α and the van der Waals radius r of the interacting atoms. Only those pairs of probe and target atoms are regarded for which the Lenard-Jones potential is negative and the resulting interaction consequently attractive. Only exceptions: should E_{lj} be positive and repulsive forces predominating, but a favourable hydrogen bonding interaction between the two atoms be abundant, E_{lj} is set to zero. Should the distance between the two atoms exceed 8Å, E_{lj} also set to zero (cut-off radius).

Equation (2.6) describes the electrostatic fraction of the interaction energy. In this equation p and q are the electrostatic charges of the target and the interacting probe, which are separated in space by distance d; K is a combination of geometric factor and natural constants. It is assumed that a planar interface is separating a homogeneous target phase of dielectricity ζ from a likewise homogeneous solvent phase of dielectricity ϵ. The nominal depth s_q of each target atom in the target phase as well as the nominal depth s_p of every single atom of the probe in solvent phase are determined by counting of all those neighbouring atoms for which the center is not further away than 4Å from the currently regarded atom. Equation (2.6) poses a compromise between the costly method of electrostatics calculation of a system according to the algorithm of Warwicker and Watson (Warwicker and Watson, 1982) and classical functions for which the shortcoming in the application to interactions with proteins was already discussed by Hopfinger (1973).

Equation (2.7) describes the fraction of the interaction energy contributed by hydrogen bonds. This directional 6-4 potential as postulated by Brooks et al. (Brooks et al., 1983) includes the tabulated constants C and D, which constitute of the hydrogen bonding parameters J of the atoms involved. If the target atom is acting as hydrogen donor, the direction of binding is determined by the position of the hydrogen atom, as emanating from the coordinates of the heavy atoms of the target. θ depicts the angle between the target donor atom to which the hydrogen is covalently bound and the probe atom acting as an acceptor. If a probe atom is acting as donor, it is assumed that the probe will orient in a way such that the most effective hydrogen bonding interaction with the acceptor can be established and $\cos(\theta)$ is set to one.

This basic concept of GRID as created in 1985 has been under continuous and consistent development since. Terms which account for the influence of a temperature factor on hydrogen bonds have been added to E_{hb} (Boobbyer et al., 1989). Further probe molecules have been added and the energy function was adapted to handle probes which can form multiple hydrogen bonds at once (Wade and Goodford, 1993; Wade et al., 1993; Wade and Goodford, 1989). Such it became possible to include water as the dominating medium in which the vast majority of natural processes takes place into the list of probes. This enables to account for a competitive effect between water and the probe and such include entropic terms into the energy function.

2.4 Employed scoring schemes

2.4.1 GRID based scoring schemes

The scoring scheme described here uses the GRID method as developed by P. Goodford (Goodford, 1985) (cf. section 2.3 on page 44). Up to eighteen different energy contour surfaces accounting for the binding properties of various small molecular fragments (as typically presented on a protein surface) are calculated for each of the complex partners. This is done by empirically parameterised physical potentials specially designed to represent binding properties of protein molecules. Besides the fifteen protein-like small chemical probes, three solvents were selected in order to allow the calculation of solvent effects emerging from the competition of atoms of the binding partner with solvent molecules. For a list of the selected probes see table 2.3 on the following page. All atoms have previously been labelled according to one of the 40 atom groups as proposed by Melo and Feytmans (1997) (see figure 2.1). Table 2.4 yields correlations, i.e. matching properties between the selected probes and the atom groups. According to this, one or more atom group number(s) have been assigned to each of the probes as defined by GRID.

Each complex conformation -as proposed by a rigid-body FFT docking algorithm- is subsequently evaluated by summation of those energy values where an atom matching the properties and requirements of the respective energy function is found in close proximity. This is done for each of the eighteen specialised force-fields used. Additionally the solvent effects for water, a hydrophilic and an amphiphilic solvent are calculated whenever atoms of the binding partner would displace or replace a solvent molecule. This yields up to 21 different scoring schemes.

In detail, the developed scoring scheme sums up the energy for every probe used to describe a protein's surface energetically - and thereby its possible binding preferences for another protein ligand. Basically, the GRID energy values for all the 18 different probes are calculated for both receptor and ligand of a complex using a grid spacing of 1Å.

The actual program sets up a score value for every probe and in addition three further scoring factors that account for solvent effects, based on the energy values for the probes defining water, an amphiphatic and a hydrophobic probe. Each complex conformation

Table 2.3: List of selected protein-like probes together with their directive symbols as used by GRID.

Probe	GRID-directive	chemical characterisation	corresponds in protein
		Single atom probes (12)	
$-CH_3$	**C3**	methyl group	(A, V, L, I, M)
$=CH-$	**C1=**	aromatic or vinyl methine group	(F, W, H, Y)
$-NH-$	**N1**	neutral flat NH group e.g. amide	(backbone, W, H)
$=NH^+$	**N1=**	sp2-hybridised imine cation	(H)
$-NH_2$	**N2**	neutral flat NH2 group e.g. amide	(N, Q, R)
$=NH_2^+$	**N2=**	sp2 hybridised imine cation	(R)
$-NH_3^+$	**N3+**	sp3 hybridised amine cation	(K, N-Terminus)
$-OH$	**O1**	alkyl hydroxy group	(S, T)
$-OH$	**OH**	phenyl or carboxyl hydroxy group	(D, E, N, Q, Y)
$=O$	**O**	sp2 hybridised carbonyl oxygen	(backbone)
$=O$	**O::**	sp2 hybridised carboxyl oxygen	(D, E, N, Q)
$-O^-$	**O-**	sp2 hybridised phenolate oxygen anion	(Y)
		Multi atom probes (3)	
$-COO^-$	**COO-**	aliphatic carboxylate group	(D, E, C-Terminus)
$-CONH_2$	**CONH2**	aliphatic neutral amidine group	(N, Q)
$-CN_2H_4^+$	**AMIDINE**	aliphatic cationic amidine group	(R)
		Solvent probes (3)	
H_2O	**OH2**	water as hydrophilic probe	
???	**BOTH**	amphiphatic probe (purely hypothetic)	
C_6H_6	**DRY**	benzene like hydrophobic probe	

as proposed by the algorithm of CKORDO (Zimmermann, 2002) is generated from the input structure and the individual atoms of the protein transformed during the docking process are then mapped onto the grid of the static protein. Every atom of the transformed protein is assigned to a single grid point on the static protein, which represents an energy vector in the dimensions of the number of probes used. As uncertainty/search radius for this mapping 1.6Å was chosen, representing the average van der Waals radius of the four most abundant elements in a protein (C, O, N, S). Depending on the atom group of the assigned atom, the corresponding probe(s) is/are retrieved from a lookup table similar to table 2.4. The respective energy value as computed by GRID for this point of the grid is then added to the total probe score for this conformation. Not only are the appropriate energies summed up to a total score for each probe, but also, whenever such a summation is performed, simultaneously the

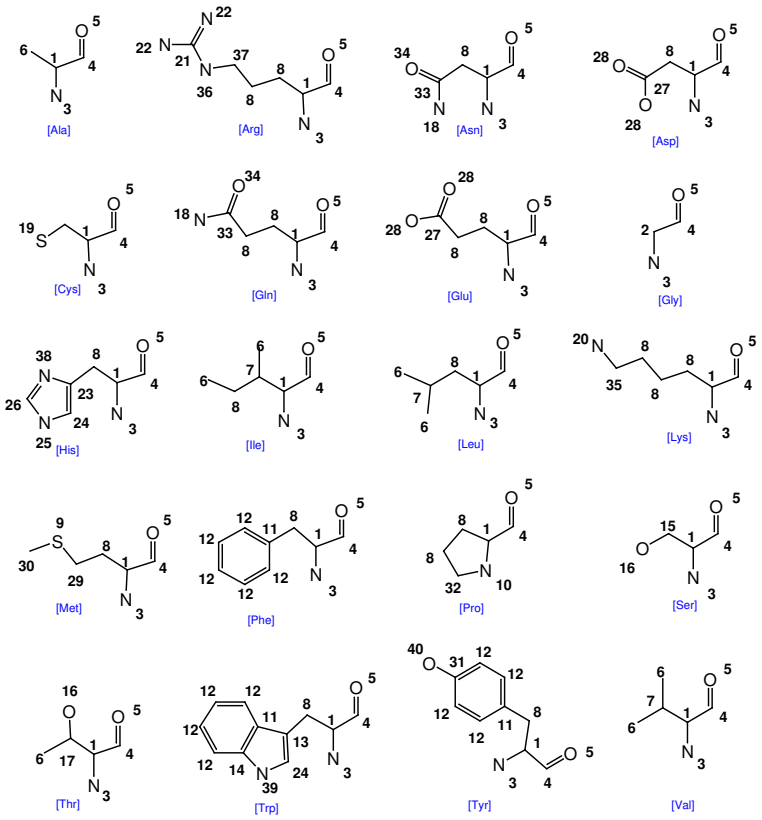


Figure 2.1: Classification of amino acid atoms in 40 atom types as proposed by Melo and Feytmans (1997).

respective energy values for the probes OH2, BOTH and DRY are added to the three score values for the solvent effects (cf. figure 2.3 on page 52). Thus, a competitive effect can be taken into account measuring the energy that will be lost or gained if the solvent at this point on the surface of the static protein is replaced by an atom of the transformed protein.

A more descriptive and schematic overview of the working procedure described above is depicted in figure 2.3 on a single step example.

GRID reportedly gives reliable energy values for unfavourable interactions only up to a value of 5kcal/mol (higher values are usually the result of clashes)(Goodford, 1985). Any occurrence of energy values above this threshold was not taken into account for the calculation of the scores described above. Whenever a corresponding atom was placed in immediate neighbourhood of such a highly unfavourable grid point, this occurrence

Figure 2.2: Schematic illustration of the scoring method: Preparation steps for the score calculation. For a given complex orientation (1) a grid is wrapped around each of the complex partners (2a: ligand, 2b: receptor). For every grid point, the energy is calculated for various small chemical probe molecules using GRID (3a,b). All grid points with a distance above a certain threshold to any of the protein atoms are discarded (4a,b). The complex partner is now placed in the respective GRID field.

Table 2.4: Correlations between atom groups according to Melo and Feytmans (1997) and probes as used by GRID.

Probe(GRID-directive)	assigned atom groups	corresponding chemical environment
C3	6	Ala, Val, Leu, Ile, Met
C1=	12	Phe, Trp, His, Tyr
N1	3	Backbone, Trp, His
N1=	38	His
N2	18	Asn, Gln, Arg
N2=	22	Arg
N3+	20	Lys, N-Terminus
O1	16	Ser, Thr
OH	28	Asp, Glu, Asn, Gln, Tyr
O	5	Backbone
O::	28	Asp, Glu, Asn, Gln
O-	40	Tyr
COO-	27, 28	Asp, Glu, C-Terminus
CONH2	18, 33, 34	Asn, Gln
AMIDINE	21, 22	Arg
OH2	10, 19, 25, 36, 39	hydrophilic, solvent
BOTH	1, 2, 4, 9, 14, 15, 17, 23, 24, 26, 29, 30, 31, 32, 35, 37	amphiphatic, solvent
DRY	7, 8, 11, 12, 13	hydrophobic, solvent

was counted. This count can optionally be used as a penalty score for the respective probe.

Since the GRID calculations can be computed in advance, the actual calculation step is extremely fast, as a very fast and efficient algorithm for approximate nearest neighbour searching based on binary kd-trees (Arya et al., 1998) was used for the grid mapping. The general working scheme of a geometric rigid body FFT based docking algorithm remains unchanged since its first development by Katchalski-Katzir et al. (1992). This involves depicting the larger unit, the receptor, as static and not to be moved in space, while the smaller unit, the ligand, is rotated and translated around the receptor. This implies, that the resulting complex conformations as proposed by the docking procedure can be generated by applying the respective transformation rules to the orientation of receptor and ligand in space that was used as a starting point. Since the receptor position is kept fixed, only the transformation rules for the ligand are needed to generate any proposed orientation.

In order to allow score calculations for both sides of the interface as seen from the viewpoint of the ligand in the field of the receptor as well as the receptor in the field of the ligand, it would normally be necessary to rotate the complete grid with the

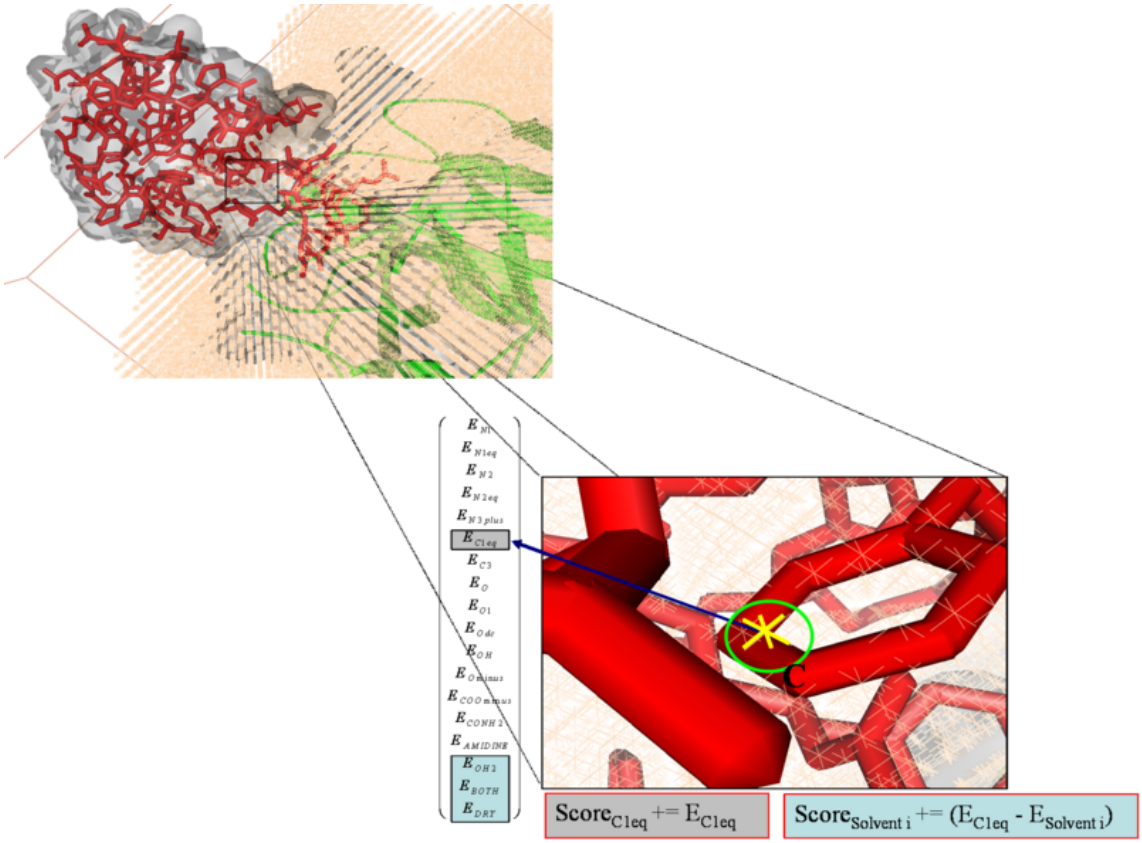


Figure 2.3: Schematic illustration of the score calculation. The picture is a two step zoom into figure 2.2, (5a) depicting the detailed working scheme of the score calculation. Whenever an atom matching the properties of the respective probe used to calculate the energy values on the grid points is found in close proximity (search radius - green) to such a point of the grid (cross - yellow), the energy value at this point is added to the respective score value. Simultaneously, the difference between this energy value to the one of any of the solvents at this point is added to the respective solvent's score.

individual force field energies for the ligand unit of a docking procedure. This is due to the fact that only transformation rules for the ligand are produced and given by the docking procedure. For each transformation which maps a vector $\vec{x}_0$ in space to a new position $\vec{x}_{trans}$ (equation (2.8)) an inverse transformation exists which will map $\vec{x}_{trans}$ back on $\vec{x}_0$ (equation (2.9)).

Transfered to docking this means that for each transformation which, if applied to the coordinates of a body L, maps L (the ligand) to a new position relative to body R (the

$$\vec{x}_{trans} = \underline{R} \cdot \vec{x}_0 + \vec{t} \tag{2.8}$$

$$\vec{x}_0 = \underline{R}^T \cdot \vec{x}_{trans} - \vec{t} \tag{2.9}$$

where:

$\vec{x}_{trans}$	:	transformed position vector in 3D space
$\vec{x}_0$	:	position vector in 3D space
$\underline{R}$	:	Rossman rotation matrix
$\underline{R}^T$	:	transposed Rossman rotation matrix
$\vec{t}$	:	translation vector

receptor), there exists an inverse transformation which, if applied to the coordinates of R, will map R to an identical relative orientation to body L as illustrated in figure 2.4. Using this mathematical relation allows for score calculation for and from the respective viewpoint of both sides of the complex interface for each individual probe/force field according to equations (2.11) - (2.12). This is computationally much more efficient, since only the atom coordinates have to be transformed (the number of atoms in a molecule will always be several magnitudes smaller than the number of grid points for the respective GRID force-field).

Each probe-specific score calculation is performed for the ligand in the immobile field of the receptor as well as the receptor in the immobile field of the ligand. The total probe specific score for a complex conformation is the sum of the scores derived for each side of the interface.

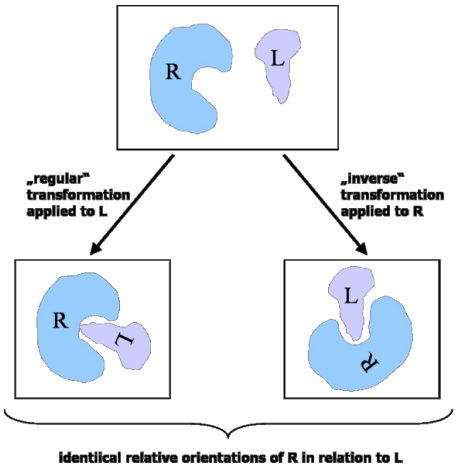


Figure 2.4: Schematic illustration of the application of "regular" and "inverse" transformations to the two binding partners of a binary protein-protein complex.

2.4.2 Residue interface propensities

Various studies (Chakrabarti and Janin, 2002; Jones et al., 2000; Lo Conte et al., 1999; Jones and Thornton, 1997), mainly of statistical nature, have been conducted in order

$$C_{probe} = \frac{1}{N_R}\sum_{i=0}^{N_R} E_{probe_i} + \frac{1}{N_L}\sum_{j=0}^{N_L} E_{probe_j} \tag{2.10}$$

$$C_{sol} = \frac{1}{N_R}\sum_{i=0}^{N_R} (E_{probe_i} - E_{sol_i}) + \frac{1}{N_L}\sum_{j=0}^{N_L} (E_{probe_j} - E_{sol_j}) \tag{2.11}$$

$$S_{probe_{pen}} = \sum_{i=0}^{N_R} i + \sum_{j=0}^{N_L} j \quad \text{if} \quad E_{probe_i}, E_{probe_j} > 5.0 \quad \frac{kcal}{mol} \tag{2.12}$$

where:

C_{probe}	:	energy cost function for probe (as taken from table 2.3 on page 48)
$E_{probe_{i,j}}$	:	energy as calculted by GRID for regarded grid point and respective probe in correlation with atoms i, j
i, j	:	interface atoms of receptor / ligand
N_R	:	total number of atoms correlating with probe (cf. table 2.4 on page 51) on receptor side of interface
N_L	:	total number of atoms correlating with probe on ligand side of interface
C_{sol}	:	energy cost function for solvent effect
$E_{sol_{i,j}}$	:	energy as calculated by GRID for regarded grid point and respective solvent probe in correlation with atoms i, j
$S_{probe_{pen}}$	:	penalty score for regarded probe

to determine, whether certain amino acids have a higher frequency of occurrence in or around the interface regions of co-crystallized complexes. Methods and results as obtained by Lo Conte et al. (1999) seemed most suitable to be integrated into the algorithm. Their proposed residue interface propensities have been derived from an analysis of the atomic structure of the recognition sites seen in 75 protein-protein complexes of known three-dimensional structure. Among these complexes were 24 protease-inhibitor, 19 antibody-antigen and 32 other complexes, including nine enzyme-inhibitor and 11 that are involved in signal transduction. The area-based composition of these 75 complexes has been analysed and was used to derive the propensities for a residue to be part of a protein-protein interface as listed in table 2.5 on the facing page, named P_{uni}. Two more interface propensity scales have been integrated which, unlike the rather universal scale for P_{uni}, have been derived from protein families and represent specialised propensities for the classes of Enzyme-Inhibitor (P_{EI}) and Antibody-Antigen complexes (P_{AA}). These family specific residue interface propensities have been derived by Huang and Schroeder (2005) from the PSIMAP database (Park et al.,

2001), which holds the contact information for more than 40,000 interfaces derived from over 8,000 PDB structures. In detail, residue interface propensities were derived from 747 interactions belonging to the SCOP (Andreeva et al., 2004) family b41.1.2 (trypsin-like serine proteases) and 612 interactions which are classified according to family b1.1.2 (C1 set domains; antibody constant domain like) using a simple mole-fraction method. In order to make the three different propensity scales comparable as well as to put an emphasis on those residues with high interface propensities in the calculation of an average residue propensity score for the interface region, the natural logarithm of the mole-fraction values was calculated as displayed in table 2.5.

Table 2.5: Propensity for a residue to be part of a protein-protein interface according to Lo Conte et al. (1999)[1] and Huang and Schroeder (2005)[2,3]

aminoacid	P_{uni}[1]	P_{EI}[2]	P_{AA}[3]	aminoacid	P_{uni}[1]	P_{EI}[2]	P_{AA}[3]
ALA	-0.43	-0.63	0.15	LEU	0.29	0.05	-0.26
ARG	0.13	-0.33	-0.08	LYS	-0.57	-0.65	-0.29
ASN	-0.12	-0.67	-0.67	MET	0.98	-0.08	0.23
ASP	-0.31	-0.05	0.12	PHE	0.79	0.66	1.00
CYS	0.76	2.20	0.00	PRO	-0.24	-0.37	0.28
GLN	-0.36	-0.45	-0.17	SER	-0.42	0.22	-0.29
GLU	-0.47	-0.60	-0.17	THR	-0.36	-0.06	-0.49
GLY	0.02	-0.03	-0.37	TRP	1.24	1.43	0.68
HIS	0.64	0.72	0.01	TYR	1.05	0.41	1.27
ILE	0.56	-0.12	-0.15	VAL	0.08	0.07	-0.09

All those residues were defined as interface residues for which at least one of their atoms is within an euclidean distance of 6Å to any atom of a respective residue of the interaction partner. For every interface residue, the respective interface propensity is retrieved and the mean value for the complete interface is calculated according to equation (2.13). Since three different propensity scales can be used, three different scoring schemes based on residue interface propensities (S_{RIP}) have been set up; one which should be generally applicable ($S_{RIP_{uni}}$), a second which is specialised for Enzyme-Inhibitor complexes ($S_{RIP_{EI}}$) and a third one which focuses on Antibody-Antigen complexes ($S_{RIP_{AA}}$).

2.4.3 Residue-residue pair potential

In order to extend the current version of the docking algorithm CKORDO, an empirical residue level pair potential has been added to the list of scoring schemes used in this

$$S_{RIP_{UNI/EI/AA}} = \frac{1}{N_R}\sum_{i=0}^{N_R} P_i + \frac{1}{N_L}\sum_{j=0}^{N_L} P_j \tag{2.13}$$

where:

$S_{RIP_{UNI/EI/AA}}$	:	Residue interface propensity score
i, j	:	interface residue of receptor / ligand
N_R	:	total number of residues on receptor side of interface
N_L	:	total number of residues on ligand side of interface
$P_{i,j}$	:	interface propensity for regarded residue i,j according to table 2.5

work. The choice fell upon the method developed by Moont et al. (1999) named RPSCORE. This residue-residue potential was derived from 103 non-homologous interfaces found in the PDB via the aid of SCOP, version 1.53. The individual scores for the possible residue pairings have been calculated using a mole fraction method according to equation (2.14) - (2.18) from a total of 10,929 residue pairings issuing from 32,439 interface residues. A pair of interface residues was defined as contacting if the distance of any of the respective atoms was below a distance cutoff of 4.5Å.

This allows for a 20x20 scoring matrix to be set up. The final value for this scoring function is calculated as the sum of the individual scores s_{ij} for a residue contact pair of types i,j within the distance cutoff of 4.5Å according to equation (2.19).

2.4.4 Tightness of fit

The *tightness of fit* scoring scheme as proposed by Gottschalk et al. (2004) is based on a normalised average minimum distance of the predicted interfacial C-alpha atoms of a protein to any of the C-alpha atoms of the binding partner. It can be calculated according to equation (2.20). A C-alpha atom was counted as a predicted interface C-alpha atom if the exponential value of the interface propensity of a residue P_i as given in table 2.5 reaches a minimum value of 1.5. Since three different scales for the residue interface propensities are employed, three different scores (ToF_{uni}, ToF_{EI} and ToF_{AA}) can be calculated.

$$s_{ij} = s_{ji} = log\frac{c_{ij}}{e_{ij}} \tag{2.14}$$

$$c_i = \sum_{j=1}^{j=20} c_{ij} \tag{2.15}$$

$$C = \sum_{i=1}^{i=20} c_i \tag{2.16}$$

$$N = \sum_{i=1}^{i=20} n_i \tag{2.17}$$

$$e_{ij} = C \cdot \frac{n_i}{N} \cdot \frac{n_j}{N} \tag{2.18}$$

where:

i, j	:	interface residue of type i, j
s_{ij}	:	score value for residue pairing between i and j
$n_{i,j}$	:	total occurrences of residues i, j
N	:	: total number of residues
c_i	:	occurrence of residues i and in contact pair c_{ij}
C	:	total number of occurrences of residues in contact pairs
e_{ij}	:	expected number of pairs between residues i and j according to mole fraction method

$$S_{RPscore} = \sum_{i=0}^{20} \sum_{j=0}^{20} s_{ij} c_{ij} \tag{2.19}$$

where:

$S_{RPscore}$	:	residue potential score
i, j	:	interface residue of type i,j
s_{ij}	:	score value for residue pairing between i and j
c_{ij}	:	total occurrences of residues pairing ij

2.4.5 Atom-atom pair potential

In order to judge the probability whether or not the distribution of atomic contacts in a proposed complex conformation is close to the native one, Grimm (2003) developed an empirical atom-atom pair potential. This potential is based on the distance dependent statistical evaluation of atom-atom contacts in protein-protein complexes. Atoms are classified into 40 atom types (see 2.1 on page 49) and contacts up to a maximal distance

$$ToF = \frac{d_{inter} - d_{all}}{d_{all}} \tag{2.20}$$

$$d_{inter} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_{inter_i}}{P_i} \tag{2.21}$$

$$d_{all} = \frac{1}{m} \sum_{j=1}^{m} \frac{D_{all_j}}{P_j} \tag{2.22}$$

where:

ToF	:	tightness of fit at the predicted binding site normalised by the size of the protein
d_{inter}	:	average minimum distance of the predicted interfacial C_α atoms of protein 1 to any of the C_α atoms of the binding partner
d_{all}	:	average minimum distance of all C_α atoms of protein 1 to any of the C_α atoms of the binding partner
D_{inter_i}	:	minimum distance of the C_α of residue i, predicted to be interface, of protein 1 to any C_α of protein 2
D_{all_j}	:	minimum distance of the C_α atom of surface residue j, which is either interface or not, of protein 1 to any C_α atom of protein 2
n	:	number of predicted interfacial residues (threshold: $P_i \geq 1.5$)
m	:	total number of surface residues
$P_{i,j}$	:	exponential value of residue interface propensity as given in table 2.5

of 8Å partitioned in 23 different distance bins. A trapeze function is used to smooth the discrete distribution function and hydrogen bonds and contacts between functional groups are taken into account as weighting factors while a repulsive part penalises steric overlaps. The observed frequencies of occurrence are transferred into pseudo energies using an empirical function. The potential is derived from a curated non-redundant dataset consisting mainly of the COMBASE database (Vakser and Sali, 1999). This knowledge based atom pair potential is integrated in the current version of CKORDO but had to be reimplemented to allow for an examination of complex candidates which are not or cannot be created by the docking procedure (due to the nature of the Fourier transformation).

2.4.6 Atomic contact energies

Based on the work of Miyazawa and Jernigan (1985), Zhang et al. (1997) computed atomic desolvation energies for 18 different atom types (cf. figure 2.5) based on a non-

redundant (maximum 25% sequence homology), high-resolution (resolution $\leq$ 2.0Å) data set of 89 protein complexes. Two atoms are defined to be in contact if their centers are within 6 Å of each other. The normalised energy values for each possible contact pair can be stored in an 18x18 matrix. An implementation for the calculation of

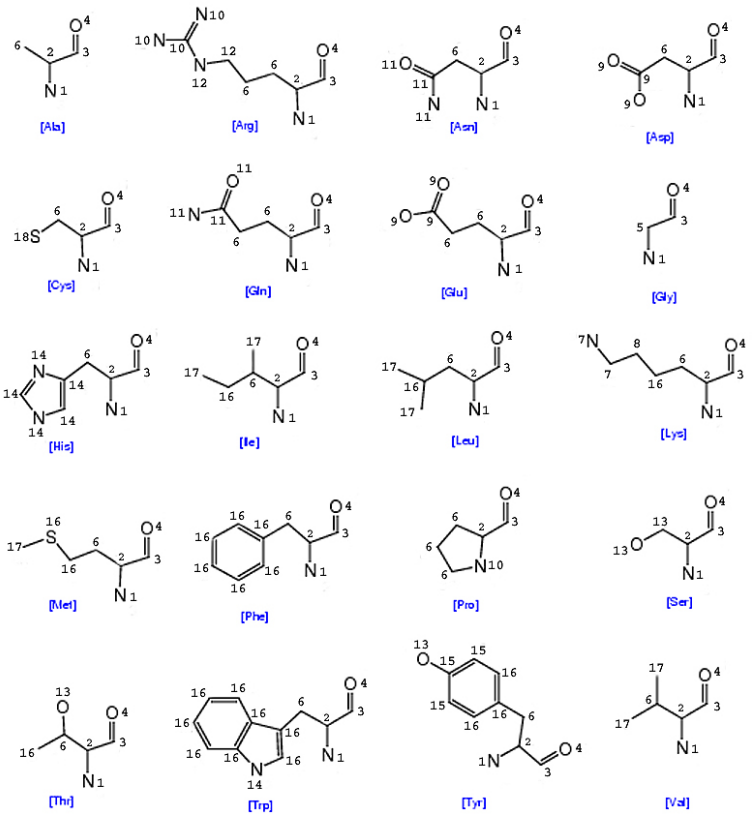


Figure 2.5: Classification of amino acid atoms in 18 atom types as proposed by Zhang et al. (1997).

these effective atomic contact energies (ACE), the desolvation free energies required to transfer atoms from water to a protein's interior, has been integrated into the software. This specialised atom-atom pair potential is known to predict the desolvation energies upon complex formation well and has already been successfully used in other docking algorithms like for example ZDOCK (Chen and Weng, 2002). In order to calculate the desolvation energy upon complex formation E_C according to the method above for an exemplary complex A-B, one has to calculate E_C for both subunits in the unbound state as well as for the complex A-B, each time considering all intramolecular atom pairs

$$E_C = \sum_{i=0}^{18} \sum_{j=0}^{18} E_{ij} n_{ij} \tag{2.23}$$

where:

E_C	:	effective contact energy of a molecule
i, j	:	atom types as seen in figure 2.5 on the page before
E_{ij}	:	effective atomic contact energy between two atoms of type i and j in contact (distance $\leq$ 6Å)
$n_{i,j}$	:	total number of atoms of type i forming contact pair with atom of type j

within 6Å. The atomic contact energy of complex formation would then be calculated according to $E_{C_{form(A-B)}} = E_C(A-B) - E_C(A) - E_C(B)$ using the formula given in equation (2.23). The current implementation uses a computationally much more efficient approach for the calculation of the desolvation energy of complex formation via the direct calculation of intermolecular atom contacts only, giving direct access to the energy term resulting from $E_C(A-B) - E_C(A) - E_C(B)$.

2.4.7 Evolutionary relationship

The degree of conservation at each amino acid site is similar to the inverse of the site's rate of evolution; slowly evolving sites are evolutionarily conserved, while rapidly evolving sites are variable. With respect to the evolution of protein-protein interactions being optimised for functional efficacy, this concept can be used to possibly distinguish native from non-native interaction sites. The method used to quantify the evolutionary relationship in this work is the one established by Glaser et al. (2003). This method, accessible via the ConSurf web-server or the standalone program Rate4Site, extracts the sequence from the PDB structure data file and automatically carries out a search for close homologous sequences of the protein of known structure. It then multiply aligns the sequences, builds a phylogenetic tree consistent with the multiple sequence alignment, and calculates the conservation scores using a Maximum Likelihood approach. In detail, the PSI-BLAST (Altschul et al., 1997) heuristic algorithm with default parameters is used to collect homologous sequences from the Swiss-Prot database (Boeckmann et al., 2003) via a single interaction of PSI-BLAST with an E-value cutoff of 0.001. The E-value or expectation value is a parameter describing the number of hits one can expect by chance when searching a database of a

particular size. The higher the E-value, the more hits will be expected, but the pairwise distance between them and the query sequence will increase. The minimum number of unique sequences required is set to five while a maximum of 50 unique sequences is used for further steps - falling back to those 50 with the highest E-value if more unique sequences could be detected. The multiple sequence alignment of the homologues extracted from the PSI-BLAST output is performed by CLUSTAL W (Thompson et al., 1994) using default parameters. The program constructs evolutionary trees consistent with the resulting multiple sequence alignment and calculates the rate of evolution at each site using the maximum likelihood paradigm (Pupko et al., 2002). This allows taking into account the stochastic process underlying sequence evolution within protein families and the phylogenetic tree of the proteins in the family. The conservation score at a site corresponds to the site's evolutionary rate. The conservation scores are normalised to a mean of zero and a standard deviation of one. The scoring function calculates the sum of the mean values for the interfaces of receptor and ligand of the proposed complex structure according to equation (2.24), giving a conservation index in the interacting region. The relevant criterion for a residue to be part of the interface is a euclidean distance of less than 6Å of any of its atoms to an atom of the interaction partner.

$$S_{Cons} = \frac{1}{N_R}\sum_{i=0}^{N_R} c_i + \frac{1}{N_L}\sum_{j=0}^{N_L} c_j \tag{2.24}$$

where:

S_{Cons}	:	conservation score (average conservation index)
i, j	:	interface residue of receptor / ligand
N_R	:	total number of residues on receptor side of interface
N_L	:	total number of residues on ligand side of interface
$c_{i,j}$	:	conservation score for regarded residue i,j

A recent quantitative analysis of interfacial amino acid conservation in protein-protein hetero complexes (Reddy and Kaznessis, 2005) indicates that the average conservation index of interface patches is not necessarily higher compared with other surface regions of the protein structures. Instead, the study reveals that the surface density of highly conserved positions is significantly higher in interface regions of protein-protein complexes which do not belong to the class of Antibody-Antigen complexes. This

is to be expected since the variable region of the antibody represents the interacting region with the antigen. These findings demonstrated that the number of conserved residues in the interacting region is a potentially more appropriate indicator for the prediction of protein-protein binding sites in most cases, while the number of non-conserved residues would be an appropriate indicator for binding-site prediction in the case of Antibody-Antigen complexes. In this work the number of highly conserved interface residues as well as the number of highly variable residues was counted for every complex conformation as proposed by the docking algorithm, where residues with a conservation score $c \leq -0.65$ were defined as highly conserved and those with conservation scores $c \geq 0.65$ defined as highly variable.

2.4.8 Temperature factors

Reportedly, interface sites tend to have lower B-Factors already in the unbound state (Yuan et al., 2003; Neuvirth et al., 2004). This easily accessible criterion has, along with others, successfully been applied to re-rank docking solutions (Gottschalk et al., 2004). Since the B-Factors, as given for a protein structure in the PDB-file are experimentally determined values which are not calculated to a standardised reference state, these values are prone to outliers and require normalisation. A median based method to detect outliers (Smith et al., 2003) was used. Therefore, the median of the B-factors in a molecule was calculated and the median of absolute displacements (MAD) determined. This allows for a so called M-value to be calculated for each B-Factor according to equation (2.25). An M_i value of ≥ 3.5 was used to define an outlier. After removal of the outliers, the remaining B-factors were normalised using Z-scores (2.26) such that the normalised B-factors have a zero mean value and unit variance. The current implementation calculates the mean value for the normalised temperature factors (after removal of outliers) of all interface atoms S_{TF} according to equation (2.27).

2.4.9 Approximation of the buried surface area

The buried surface area can be defined as the area of the protein surface of a complex subunit that is freely accessible to solvent molecules in the unbound state while becoming inaccessible to solvent upon complex formation. Usually the buried surface area is calculated as the sum of the solvent accessible surface areas of the complex

$$M_i = 0.6745 \cdot \frac{(x_i - \tilde{x})}{MAD} \tag{2.25}$$

$$z_i = \frac{x_i - \bar{x}}{\sigma} \tag{2.26}$$

where:

M_i	:	M-value at atom position i
x_i	:	B-Factor at atom position i
$\tilde{x}$	:	median of the B-factors
MAD	:	median of absolute displacements (absolute displacement: $x_i - \tilde{x}$)
z_i	:	Z-score for measured value x_i
$\bar{x}$	:	mean value for all x_i
σ	:	standard deviation for all x_i

$$S_{TF} = \frac{1}{N_R} \sum_{i=0}^{N_R} b_i + \frac{1}{N_L} \sum_{j=0}^{N_L} b_j \tag{2.27}$$

where:

i, j	:	interface atom of receptor / ligand
N_R	:	total number of atoms on receptor side of interface
N_L	:	total number of atoms on ligand side of interface
$b_{i,j}$	:	normalised B-factor for regarded atom i,j

components minus the solvent accessible surface area of the complex. The calculation of the solvent accessible surface however is non-trivial since it affords a complete mathematical description of the protein surface, e.g. via triangulation. The current version of the CKORDO program avoids this with the help of the time consuming call of the external program DSSP for every proposed conformation. As an alternative to this usage of an external program, a simple but effective method to approximate the buried surface area has been implemented which allows to completely dispense with the calculation of solvent accessibilities for proposed complex conformations.

The atomic solvent accessibilities are precomputed once only for the subunit structures using NACCESS (Hubbard and Thornton, 1993a) and stored in a modified PDB format which has been used as a standard for all the calculations in this work. In order to be solvent accessible, there is the need for sufficient space in the proximity around a regarded atom that can be occupied by a solvent molecule, usually water, without steric hindrance. Using a simplified representation of a water molecule as a sphere with a

radius of 1.4Å leads to the approximation that only those atoms will contribute to the buried surface area with their solvent accessibility assigned in the unbound state, which do not allow a sphere of 2.8Å diameter to be fit in between their van der Waals surfaces. Assuming a maximal van der Waals radius of 1.9Å for a protein atom, the upper limit for the distance of the centres of two atoms forming an inter molecular atom contact pair across the complex interface equals 6.6Å. All inter molecular atom contact pairs above this distance threshold do not contribute to the approximated buried surface area. Since this method is not able to detect cavities which are large enough to contain solvent molecules but are still occluded from the solvent since the surrounding areas are in close contact with an interaction partner, it rather resembles a contact surface than the true buried surface area. The final score for the approximated buried surface area is computed according to equation (2.28).

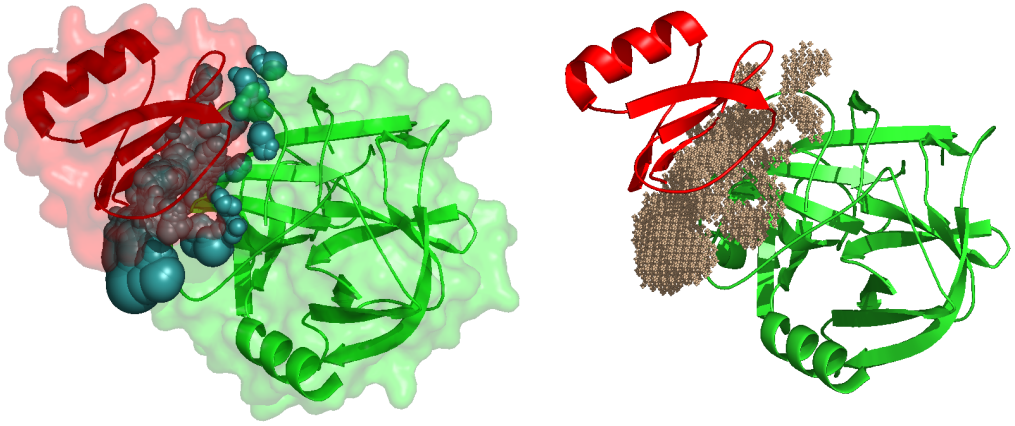
$$S_{bursurf} = \frac{1}{2}\left(\sum_{i=0}^{N_R} asa(i_{ref}) + \sum_{j=0}^{N_L} asa(j_{ref})\right) \tag{2.28}$$

where:

$S_{bursurf}$	:	score value for approximated buried surface area (contact surface)
i, j	:	interface atom in distance $d_c < 6.6$Å and with surface distance $d_s > 2.8$Å
N_R, N_L	:	total number of interface atoms of receptor and ligand, respectively
$asa(i_{ref}, j_{ref})$	:	accessible surface area of atoms i or j in unbound reference state

2.4.10 Calculation of the gap volume

CKORDO offers the possibility to calculate the volume in between the interacting subunits. This is done via the external program SURFNET (Laskowski, 1995) and the subsequent parsing of the relevant output. Since such a procedure is time consuming, an algorithm for the computation of the gap volume for each complex conformation as proposed by a docking algorithm has been implemented for this work.
The gap volume definition implemented is based on the SURFNET methodology through which the gap regions are built up by fitting of spheres into the spaces between atoms, considering all relevant pairs of atoms in turn and placing a sphere midway in between each pair, reducing its size if it clashes with any neighbouring atom.



(a) final set of gap spheres **(b)** gap volume flooded by grid

Figure 2.6: Illustration of the gap volume calculation method shows native conformation of enzyme inhibitor complex between alpha-chymotrypsin with eglin C from oxen (*Bos taurus*) and leech (*Hirudo medicinalis*) (PDB-ID 1ACB): (a) molecules shown as cartoon representation with translucent van der Waals surface (enzyme in green, inhibitor in red) including the final set of gap spheres (blue) (b) molecules shown as cartoon representation with centres of grid cubes used to flood volume captured by spheres from (a) shown as crosses.

In detail, all intermolecular atom pairs in contact distance $d = 2 \cdot r_{sphere_{max}} + r_{vdW_{max}}$ have to be retrieved, where $r_{sphere_{max}}$ is the maximal radius of the initially placed gap sphere and $r_{vdW_{max}}$ the maximal van der Waals radius for any atom of the molecules. In the case of proteins (Phosphorus is excluded in calculations), $r_{vdW_{max}}$ equals the van der Waals radius of Carbon and is set to 1.87Å. The maximal radius for initial gap spheres is set to 4.0Å as default, allowing for a maximal distance of two interface atoms' surfaces of 8Å. By the initial placement of "trial spheres" of a defined maximal size between every contacting atom pair of the subunits, the problem of setting distinct boundaries for the posterior volume calculation can be handled. Subsequently, the radii of the initial spheres are then reduced whenever any neighbouring atoms are found to penetrate it until all neighbouring atoms of inter- and intramolecular nature have been considered and one is left with a final, possibly shrunken gap sphere. If the sphere is still above some minimum size (default of 1.0Å), the position for its centre is stored along with its radius. This procedure is repeated until all possible pairs of atoms have

been considered and one is left with a set of gap spheres filling the region between the molecules (see figure 2.6 on the preceding page, (a)). Since the resulting final set of gap spheres is highly overlapping, the actual calculation of the volume captured by the spheres is not performed analytically due to the complexity of the problem, but rather an iterative approach is chosen. Hereby, the final set of gap spheres is placed in a three dimensional grid of defined spacing and the total volume captured inside the spheres is "flooded" with individual cubes of the grid (see figure 2.6 on the page before, (b)). The accuracy of this approach can be adjusted via the chosen grid spacing (a default value of 1Å yielded reliable results).

2.5 Comprehensive scoring of protein-protein docking solutions

2.5.1 Theoretical approaches to the merging (combination and parameterisation) of individual postfilter scoring schemes

The methods described in the previous chapter 2.4 offer a potentially large number of individually calculated score values. These scores need to be combined in a sensible way with the objective of an optimal discrimination of near-native and inacceptable complex conformations as proposed by a docking algorithm. Ideally this is achieved by the creation of a single scoring scheme which allows for a fast and efficient re-ranking of docking primary results. This is generally possible in a variety of ways, three of which will be explained in detail in the following subsections.

2.5.1.1 Consecutive application of the individual scores

One possibility for the combination of individual scores is the consecutive application of the individual scores, where each score basically works as an independent ranking scheme. Only the top ranking scores are passed on to the next ranking scheme and so on. This method bears the risk of actually loosing true positive solutions at an early stage, so that even the best possible filtering methods at stages that are down the chain have no chance of retrieving these near-native solutions. However, since each step will filter out a certain number of proposed complex structures, the search space is gradually

reduced, thus bearing the possibility of saving computational resources which either could make such a scoring scheme faster and might even allow for computationally intensive filtering and scoring steps to be applied at the end of the chain. In order to achieve optimal results for a consecutive application of individual scoring schemes, decision/classification trees or recursive partitioning (Susnow and Dixon, 2003) can be applied.

2.5.1.2 Combined application of the individual scores

When using the combined application of the individual scores, a total score is generated which holds the information from all the individual scores in a consensus manner. The simple, straightforward way to do this is to produce a comprehensive scoring function via linear combination of the individual parameters (equation (2.29)). The advantage

$$S_{tot} = \sum_{i=0}^{N} \alpha_i \cdot s_i \tag{2.29}$$

where:

S_{tot}	:	comprehensive score
s_i	:	individual score
N	:	number of individual scores s_i
α_i	:	linear coefficient

of such a linear combination of scores is that it is easily extensible. Parameterisation can be achieved via a simple linear regression by optimisation of the linear coefficients. The large disadvantage of such a method is the implicit assumption that the individual scores are linear independent, which will most likely not be the case (but might still be a reasonable approximation). Optimisation would be driven such that either the docking results will be optimised for the true native or nearest-native structure to be ranked on position one or at least close for as many examples as possible or, such that the number of true positive, near-native structures (e.g. those with an interface-RMSD of less than 4Å) is enriched in the depicted area while the number of false positive solutions will have to be decreased.

2.5.1.3 Using machine learning methods to combine the individual scores

In order to combine multiple individual scores for a simultaneous application of all these filtering aspects, a machine learning approach can be used. Neuronal Networks or Support Vector Machines (Schölkopf et al., 2003; Burges, 2002; Mangasarian, 2001; Vert, 2001) can be trained on classification and ranking problems such as the one on hand. The clear advantage of such a method would be that in contrary to the approach described in the previous section 2.5.1.2 no restrictive assumption of independency of the individual variables has to be assumed, while leading to a single step comprehensive scoring function.
Due to the high number of possible scoring criteria and obvious dependencies, especially among the GRID based cost functions, the method of choice for this work involves a machine learning approach, based on a classification of putative complex conformations as proposed by a docking algorithm (see sections 2.6 and 2.8).

2.6 Classification of docking results

The scoring schemes implemented could now be optimised and combined to a "classical" scoring function for the evaluation and re-ranking of complex conformations as proposed by a docking procedure (Fernandez-Recio et al., 2004), aiming for a regression of the major quality criterion, the RMSD. An alternative approach is the classification of docking solutions into acceptable, true solutions and unacceptable, false solutions on the basis of these scoring schemes. This is sensible since a direct correlation between the root mean square deviation (RMSD) of a proposed orientation to the native structure is only useful for those solutions which can be depicted as near-native. A docking solution with an RMSD of 1Å to the native structure is more desirable than a solution with an RMSD of 5Å to the native structure, while solutions of 6Å RMSD or more can be considered as inacceptable, no matter whether the actual derivation to the native structure sums up to 10, 20 or 50Å.
Usually solutions with an RMSD larger than 4-5Å to the native complex are regarded as false solutions (Halperin et al., 2002). This criterion can be complemented by the rate of correctly matched residue pairings as compared to the native state (Janin et al., 2003). In this work a borderline is drawn at 5Å RMSD of interface C-alpha atoms to

the fitted complex, because it is highly probable that these solutions will end up closely to the native solution after a final refinement step is applied. In order to distinguish near-native complex conformations from false solutions, machine learning algorithms can be trained on the classification of these cases.
Two-state classification problems for machine learning require equipartition between the two classifiers for the training data. However, rigid-body docking usually yields few acceptable solutions among a theoretically unlimited number of unacceptable ones. This problem was handled by the artificial enrichment of existing true solutions through trial-and-error application of slight, random rotational and translational movements to these conformations. Each such generated solution was carefully checked for clashes and the minimum number of required residue-residue pairings before the RMSD was finally recalculated in order to judge whether this complex orientation could be accepted as new, *artificially enriched* true solution.

2.7 Postfilter software development

Figure 2.7 depicts the general workflow of the developed protein-protein docking postfilter software. Starting from structural data of the two units of a dimeric docking case along with the respective output of a docking algorithm (CKORDO), the required additional information is assigned to the molecules hierarchies (on atom, residue, chain, model and molecule level). For each set of transformational parameters as listed in the docking output the respective complex conformation is generated and can be compared to a reference complex. This reference complex is optionally given for the case that the input data for the subunits to be docked are not or cannot be given in the same orientation in space as the original input structures for the docking procedure. The RMSD, the number of soft and hard clashes as well as the percentage rate of residue pairings as compared to the reference complex are then calculated. If the options for an artificial enrichment have been set and the respective conformation generated fulfills the necessary conditions for an acceptable docking solution, random changes are applied to the transformational parameters used for its generation in order to create a new, similar but not identical complex conformation which again is evaluated for being an acceptable or inacceptable docking solution. This process is repeated until the required amount of artificially enriched near native complex conformations has

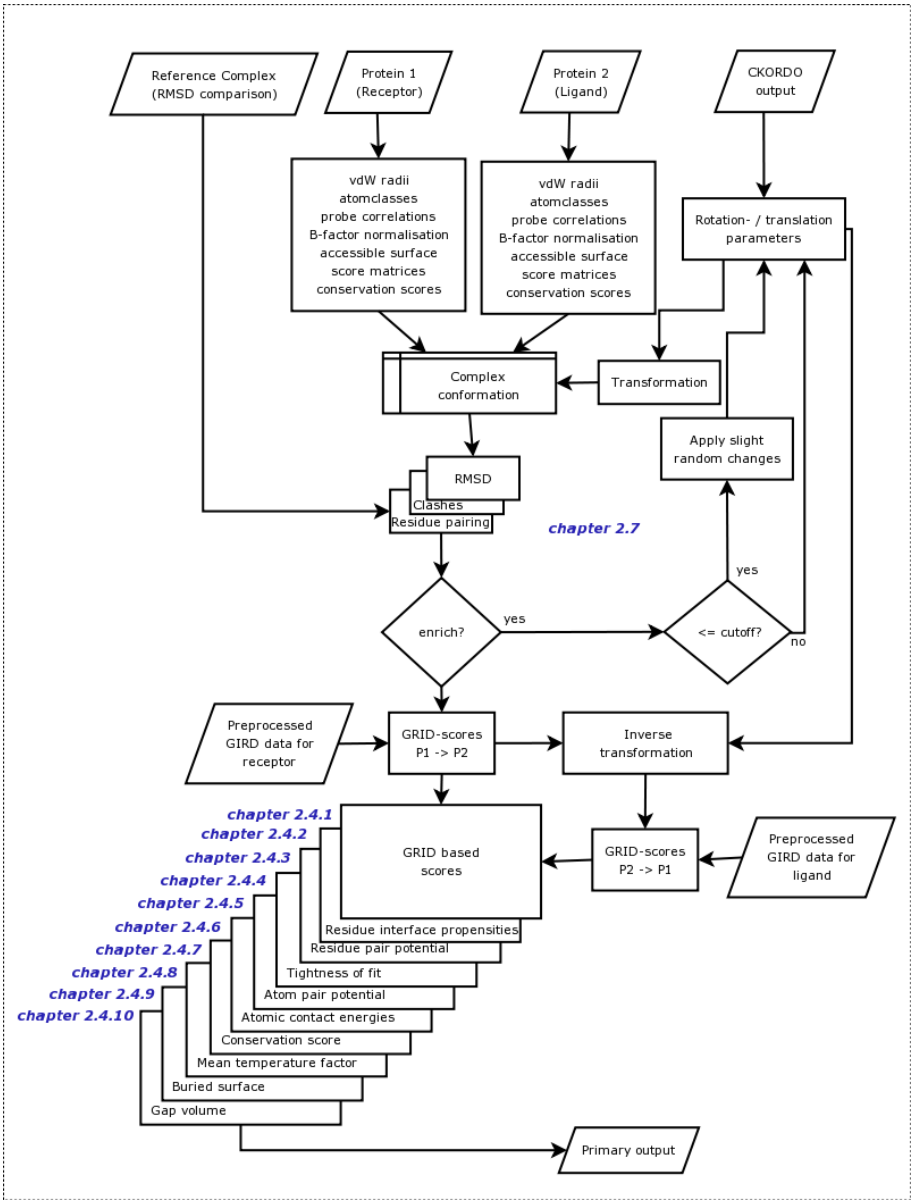


Figure 2.7: Flowchart of the developed protein-protein docking postfilter software. The individual scoring functions are described in detail in the chapters listed in the scheme.

been reached. For each of the generated complex conformations - the ones as listed in the docking output as well as the generated artificially enriched ones - the scoring values as described in chapter 2.4 are consecutively calculated.
The software is written in the C++ programming language and consists of more than 15,000 lines of code. It is intensively documented using the DOXYGEN documentation system providing availability of the documentation in a number of output formats including HTML, Man pages, RTF, XML and PDF. The handling of protein molecules, from the reading/writing of various input/output formats as provided by the PDB to the browsing and manipulation of molecular hierarchies is based on the CCP4 COORDINATE LIBRARY (Krissinel et al., 2004). The time critical procedure for the matching of GRID energy points with the atoms of the protein is accomplished using the ANN library for approximate nearest neighbour searching (Arya et al., 1998) (see also section 2.4.1 on page 47).

2.8 Machine learning

The field of machine learning studies the design of computer programs able to induce patterns, regularities, or rules from past experiences. The learner (a computer program) processes data representing past experiences and tries to either develop an appropriate response to future data, or describe in some meaningful way the data seen (Alpaydin, 2004). One can generally distinguish three different types of machine learning:

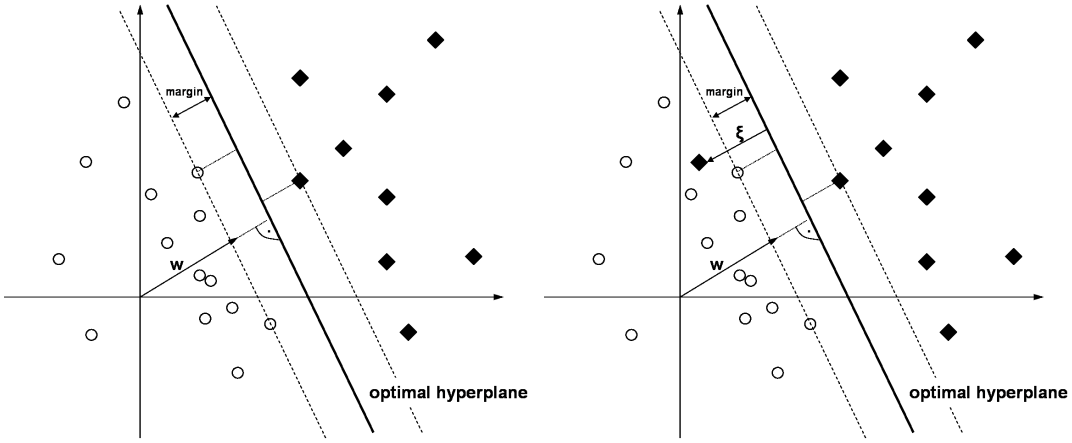
- Supervised learning
  Learning a mapping between an input x and a desired output y
- Unsupervised learning
  Understanding the relationships between data components
- Reinforcement learning
  Learning to act in the environment based on the delayed rewards

The machine learning methods applied in this work are algorithms for supervised machine learning. Supervised learning is a machine learning technique for creating a function from training data. The training data consist of pairs of input objects (typically vectors) and desired outputs. The output of the function can be a continuous

value (called regression), or can predict a class label of the input object (called classification). The task of the supervised learner is to predict the value of the function for any valid input object after having seen only a small number of training examples (i.e. pairs of input and target output). To achieve this, the learner has to generalise from the presented data to unseen situations in a reasonable way.
Machine learning has a wide spectrum of applications including search engines, medical diagnosis, detecting credit card fraud, stock market analysis, speech and handwriting recognition, game playing and robot locomotion. In bioinformatics, the usage of machine learning methods has become popular for a broad range of applications. Examples for this are DNA classification, prediction of gene function, subcellular localisation, extraction of biological relations via text mining and many more (cf. section 1.3.2). A general survey of the applications of machine learning methods in bioinformatics can be found in Baldi and Brunak (1998) and Bhaskar et al. (2005).

2.8.1 Support Vector Machines

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Two principles constitute the basis for the success of Support Vector Machines: the *maximal margin hyperplane* and the *kernel trick* (Boser and Vapnik, 1992). The central ideas of support vector learning will be described in the follwing by means of a very simple toy example.
Within a simple idealised binary classification problem of linear separable data, the two data classes can be separated by a series of hyperplanes. Among all hyperplanes separating the data, there exists a unique *optimal hyperplane*, distinguished by the maximum margin of separation between any data point and the hyperplane. This optimal hyperplane can be constructed by maximising the margins from the hyperplane to the nearest data point of each class (also called support vectors). All data points lying to one side of the plane would then ideally be of the same class, while those points lying to the other side of the optimal hyperplane would belong to the opposite class (see figure 2.8 (a)).
In practice, an ideal separating hyperplane may not exist (see figure 2.8 (b)), e.g. if a high noise level causes a large overlap of the classes. Calculation of the optimal hyperplane can also be expanded for the case of non-separable training sets. For the



(a) linear separable data **(b)** linear inseparable data

Figure 2.8: A binary classification toy problem: separate balls from diamonds. The optimal hyperplane is shown as a solid line and the maximal margin is drawn in. For the linear separable case (a), the optimal hyperplane clearly distinguishes balls from diamonds. For the linear inseparable case (b), the optimal hyperplane has to allow for the possibilities of examples violating by introduction of slack variables ξ.

example above this would mean to use linear separation while admitting training errors. Training errors are covered by introduction of an error penalty value, expressed by the distance of the erroneous data instance to the hyperplane multiplied by an error cost. This approach is called the soft margin hyperplane.

The second basic principle of SVMs, the so called *kernel trick* allows for the mapping of the data from the input space into an adequate higher dimensional space called *feature space* in which the separation of the data via an optimal hyperplane may become substantially easier. This is illustrated for another toy example in figure 2.9 where a linearly inseparable classification in input space (left hand side of the figure, data only separable by elliptical function) becomes linearly separable after mapping to a higher dimensional feature space. This approach becomes feasible since the mapping does not have to be carried out explicitly. For the calculation of the optimal hyperplane only the dot product of two feature vectors has to be calculated which is given via the application of a so called *kernel function* directly on the input data.

Mathematically the working principle of a binary classification SVM can be expressed as following: Given a training set of instance-label pairs (x_i, y_i) with $i = 1, \ldots, l$ where

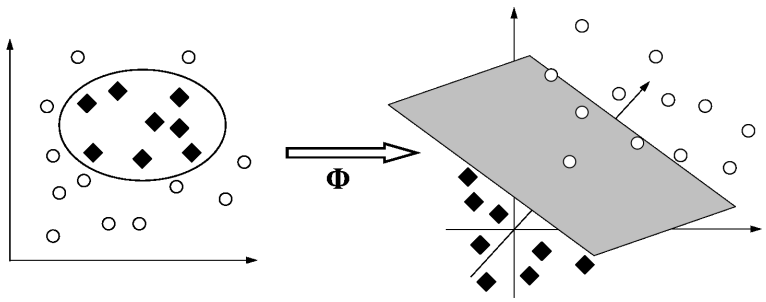


Figure 2.9: The *kernel trick* as basic idea of SVMs: map the training data into a higher dimensional feature space via a kernel function Φ and construct a separating hyperplane with maximum margin. The example on the left side involves a nonlinear decision boundary in input space to separate balls from diamonds. This complex problem in low dimension may become simpler in higher dimensions as shown on the right hand side of the picture. By the use of a kernel function, it is possible to compute the separating hyperplane without explicitly carrying out the map into feature space.

$x_i \in R^n$ and $y \in \{+1, -1\}^l$, there exists a weight vector w and a threshold b such that $y_i \left(\langle w, x_i \rangle + b\right) > 0$ $(i = 1, \ldots, l)$. Rescaling w and b such that the point(s) closest to the hyperplane satisfy $|\langle w, x_i \rangle + b| = 1$, a canonical form (w, b) of the hyperplane can be obtained, satisfying $y_i \left(\langle w, x_i \rangle + b\right) \geq 1$ respectively $y_i \left(\langle w, x_i \rangle + b\right) \geq 1 - \xi_i$ in the case of a soft margin hyperplane. To construct the optimal hyperplane, the following optimisation problem has to be solved:

$$\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{2.30}$$

$$\text{subject to} \quad y_i \left(w^T \Phi(x_i) + b\right) \geq 1 - \xi_i \quad \text{for all} \quad i = 1, \ldots, l; \quad \xi_i \geq 0; \quad C > 0. \tag{2.31}$$

where:

w	:	N-dimensional vector
w^T	:	transposed N-dimensional vector
C	:	penalty parameter of error term
ξ_i	:	slack variable accounting for training error
Φ	:	kernel function

This represents a quadratic programming problem (equation (2.30)) which can be

solved efficiently and globally using the respective constraints (equation (2.31)).
The most common basic kernel functions are listed in equations (2.33) - (2.35).

$$\begin{aligned}
&\text{linear:} && K(x_i, x_j) = x_i^T x_j && && (2.32)\\
&\text{polynomial:} && K(x_i, x_j) = \left(\gamma x_i^T x_j + r\right)^d && , \quad \gamma > 0 && (2.33)\\
&\text{radial basis function (RBF):} && K(x_i, x_j) = \exp\left(-\gamma ||x_i - x_j||^2\right) && , \quad \gamma > 0 && (2.34)\\
&\text{sigmoidal:} && K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j + r\right) && && (2.35)
\end{aligned}$$

where:
γ, d, r : kernel parameters

This work uses Support Vector Machines in order to distinguish near-native orientations of protein-protein interfaces as proposed by a docking algorithm from non-native ones (cf. section 2.6 on page 68).
All but the linear kernel from equations (2.33) - (2.35) are dependent of at least two *kernel parameters*. To get good generalisation ability, a validation procedure is conducted to decide parameters: Considering a grid space of (C, γ) of defined range and spacing, for each hyperparameter pair in the search space, an X-fold cross validation on the training data set is conducted. This procedure can be iteratively repeated for various discrete values of the third possible kernel parameter r, if required by the respective kernel function, thus adding a new dimension to the grid search and optimisation problem.
Those combinations of (C, γ, r) are subsequently chosen that lead to the lowest balanced error rate in cross validation and used to create a model as the predictor.
The tool of choice in this work has been the LIBSVM library (Chang and Lin, 2005), a C++-library offering various classes, methods and state of the art tools for efficient training of SVM models as well as their application.

2.8.1.1 Probabilistic Support Vector Machines

In their standard formulation Support Vector Machines output hard decisions rather than conditional properties (Schölkopf, 1997; Smola et al., 2000). The decision function

associated with an SVM is based on the sign of the distance from the separating hyperplane (see equation (2.36).

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i K(x, x_i) \tag{2.36}$$

where:

$f(x)$	:	SVM decision function
x	:	input vector
$\{x_1, \ldots, x_n\}$	:	set of support vectors
y_i	:	class of the i-th support vector (+1 or -1 for positive and negative examples, respectively)
N	:	number of support vectors
α_i	:	weighting factor for support vector i
K	:	kernel function

However, margins can be converted into conditional probabilities in different ways (Platt, 2000) for classification problems. This can be done by mapping margins into conditional probabilities using a logistic function (equation (2.37)) parameterised by an offset B and a slope A and adjusting these according to the maximum likelihood principle, assuming a Bernoulli model for the class variable C_i.

$$P(C_i = 1|x) = \frac{1}{1 + exp(-Af(x) - B)} \tag{2.37}$$

Aim of this work is the discrimination of near-native solutions from non-native ones in the ensemble of proposed complex conformations resulting from a protein-protein docking algorithm. Applying a classification approach using a probabilistic SVM provides the framework for the creation of a continuous scoring function.

2.8.2 Performance measures for machine learning

In order to assess and finally judge the quality of a model trained by a machine learning algorithm, it is not only important that the data on which the final prediction is performed is chosen carefully, but also that the results of the prediction can be quantified in some way. Therefore, the values returned by the predictor have to be

compared to the real data values. Figure 2.10 illustrates the usual terminology used for a binary classification problem.

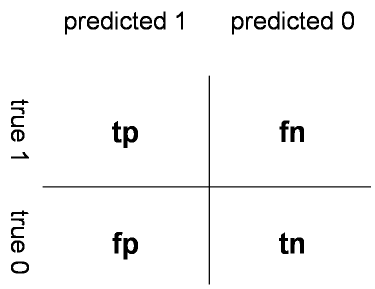


Figure 2.10: Schematic illustration of denotations used in quality measures for binary classification. A binary classification problem is assumed with a 0 value representing false and a value of 1 representing true solutions. (tp: true positive, tn: true negative, fp: false positive, fn: false negative.)

The most important since most obvious and thus most widely used quality measure for binary classification is the so called prediction accuracy (cf. equation (2.38)). Prediction accuracy is the percentage rate of correctly predicted cases, when comparing the results of a machine learning predictor to the actual real-case(s). Using prediction accuracy alone as a single quality measure can be misleading though, especially if the data that the prediction is performed on is imbalanced. E.g. if predictions would be performed on imbalanced data with very few true examples, like it is the case for a realistic docking procedure, a classifier which simply classifies every single case as false would lead to a relatively high prediction accuracy value, while at the same time being totally over-trained and such of no use at all. This simple example explains the need for further quality measures for machine learning. For a regression problem, a widely used quality measure is the correlation coefficient. Such a correlation coefficient is, in its original definition, not applicable to binary classification problems. Matthews correlation coefficient (see equation (2.39)) poses an adaption of this quality measure to two-state classifications. Accuracy and Matthews correlation coefficient can be used to judge the overall performance of a binary prediction method.

Using further quality measures, it is possible to distinguish, dissect and quantify the performance for the individual cases (true/false) further. The sensitivity or recall value (see equation (2.42)) describes the partition of correctly identified true solutions. The specificity or precision value can be defined in two ways. The common definition via positive cases describes to which amount positive/true predictions are actually correct, i.e. the reliability of a positive/true prediction (cf. equation (2.40)).

$$acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.38}$$

$$mcc = \frac{(tp \cdot tn) - (fp \cdot fn)}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{2.39}$$

$$spec^{+} = \frac{tp}{tp + fp} \tag{2.40}$$

$$spec^{-} = \frac{tn}{tn + fp} \tag{2.41}$$

$$sens = \frac{tp}{tp + fn} \tag{2.42}$$

$$f = \frac{2 \cdot (spec \cdot sens)}{spec + sens} \tag{2.43}$$

where:

acc	: prediction accuracy
mcc	: Matthews correlation coefficient
$spec^{+}$	: specificity or precision; positive prediction value (PPV)
$spec^{-}$	: specificity or precision; negative prediction value (NPV)
$sens$	: sensitivity or recall
f	: f-value, harmonic average between precision and recall

This value is also known as positive prediction value (PPV). An alternative definition via negative cases, the negative prediction value (NPV), describes the partition of correctly identified negative/false solutions (equation (2.41)). The so called f-value (equation (2.43)) represents a harmonic average between precision and recall.
Specificity/precision and sensitivity/recall should ideally equal to one. This would only be the case for an *ideal* or perfect predictor. In order to compare different predictors and their performance among each other, a plot of precision against recall, known as Receiver Operate Characteristics (ROC) is commonly used.

2.8.3 Feature selection strategies

Since the scoring schemes as described in section 2.4 will not all be of similar discrimination power, respectively will not provide strictly independent information, it might not be necessary to train machine learning algorithms on a combination of all features in order to reach the best possible results for a classification of putative

protein interactions as given by a docking algorithm.
Therefore, a feature selection approach was chosen in order to find the best possible combination of features, respectively scoring functions, for the problem at hand. Typical feature selection strategies like sequential forward- or backward selection or genetic algorithms involve numerous repeats of the actual training procedure with various feature combinations. Due to the high number of features and the amount of input (training) data instances, such an iterative feature selection procedure becomes infeasible in terms of computational effort for support vector machines using a kernel function other than the linear kernel in application to the problem. A relatively simple and fast feature selection procedure based on F-scores as presented by Chen and Lin (2004) is applied. F-score is a simple technique which measures the discrimination of two sets of real numbers:

$$F(i) = \frac{\left(\overline{x}_i^+ - \overline{x}_i\right)^2 + \left(\overline{x}_i^- - \overline{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^+ - \overline{x}_i^+\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^- - \overline{x}_i^-\right)^2} \tag{2.44}$$

where:

$F(i)$	:	F-score value for i-th feature
$x_k, k = 1, \ldots, m$	:	set of m training vectors
n_x, n_-	:	number of positive and negative training instances
$\bar{x}_i, \bar{x}_i^+, \bar{x}_i^-$	:	average of the ith feature of the whole, positive, and negative data sets, respectively
$x_{k,i}^+, x_{k,i}^-$	:	ith feature of the kth positive, and negative instance

The numerator of equation (2.44) indicates the discrimination between the positive and negative sets, and the denominator indicates the discrimination within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative. Therefore, this score can be used as a feature selection criterion.

2.9 Evaluation of scoring performance

The general aim of every scoring function for protein-protein docking is the discrimination between acceptable, near-native solutions and erroneous complex conformations. When the scoring scheme is applied, the re-ranking process should sort as many

acceptable solutions, sometimes also called hits, as possible among the top ranks of the list of possible solutions. The primary quality criterion used to distinguish acceptable, near-native solutions from erroneous ones is the root mean square deviation to either the native complex or the fitted unbound units. The quality criterion that is most often used in order to judge the performance of a scoring function is the total rank of the first truly acceptable solution - usually defined by setting an RMSD threshold - often accompanied by the absolute rank of the best possible solution; the one with the lowest RMSD. However, these listings of absolute ranks only give insight into part of the performance of a scoring function, by picking selected information. Furthermore, the total number of solutions submitted to a re-ranking procedure may vary, such that results expressed by listings of absolute ranks are hardly comparable.
This problem can be overcome by relating the number of acceptable solutions found to the number of solutions that have to be screened therefore using percentage values instead of absolute rank numberings. Plotting the percentage of acceptable solutions that have been found against the percentage of ranks that have to be searched therefore in a so called *enrichment plot* allows for a direct visual comparison between two or more rankings as they emanate from the underlying scoring schemes.
Since a realistic scoring function is hardly likely to perform perfectly, some acceptable near-native structures can be filtered out or lowered in rank together with erroneous, non-native ones in the ensemble. For the direct comparison of two different rankings, the *improvement factor* (IF) (Huang and Schroeder, 2005) can be calculated according to equation (2.45).

$$IF = \frac{N^2_{acc}/N^2_{tot}}{N^1_{acc}/N^1_{tot}} \tag{2.45}$$

where:

N^1_{acc}, N^2_{acc}	:	number of acceptable solutions (hits) according to ranking 1,2
N^1_{tot}, N^2_{tot}	:	total number of solutions according to ranking 1,2

The improvement factor is especially useful if scoring functions are applied in a consecutive manner (see section 2.5.1.1 on page 66), explicitly excluding erroneous solutions by defined threshold values from the population of putative solutions in each step. It can also be applied to a classification approach under the presumption that all docking solutions classified as inacceptable are to be excluded for further steps.

$$P = \sum_{x=a}^{b} \frac{\binom{r}{x}\binom{m-r}{n-x}}{\binom{m}{n}} \tag{2.46}$$

where:

m		:	total number of complexes subject to re-ranking (ensemble size)
n		:	rank of the first near-native (acceptable) solution; if no acceptable solution is found n is set to the rank of the best structure in the ensemble
r		:	total number of acceptable solutions in the ensemble
$a = 1$		:	at least one acceptable solution is to be found
$b = \begin{cases} r & \text{if n} > \text{r} \\ n & \text{if n} \leq \text{r} \\ 1 & \text{if r} = 0 \end{cases}$		:	upper limit of possible number of near-native structures when picking n times

A probabilistic approach to the evaluation of scoring performance is given by calculating the chance of obtaining a result as good or better than that obtained by the scoring function by randomly picking complexes out of the pool of generated complexes as proposed by Gottschalk et al. (2004). This *scoring probability* P is described by the hypergeometric distribution and can be calculated according to equation (2.46). Equation (2.46) calculates the probability of obtaining at least one near-native complex and at maximum all possible acceptable solutions by chance when picking n times. The lower this probability is, the better the scoring performance.

3 Results

"Experience does not ever err. It is only your judgment that errs in promising itself results which are not caused by your experiments."

Leonardo da Vinci, 1452 - 1519.

3.1 Primary docking and postfilter results

For all docking test cases listed in tables 2.1 on page 37 and 2.2 on page 39, CKORDO docking runs have been performed starting from the unbound units as fitted on the native complex, using a 12 degree rotational angle increment, while storing the best 5 translations for every sampled rotation step. This yielded a total of 43,080 conformations for every test case examined. The unbound docking test case for the native complex of 1N2C in table 2.2, a medium-difficulty docking case of type "Other", turned out to exceed the size limitations (in terms of total number of atoms) that the current version of CKORDO is able to handle and therefore had to be omitted.

CKORDO provides the option of using predefined rotational transformations, for which subsequently all possible translations are calculated and those of highest geometric correlation stored. Such, a much closer sampling of the conformational space can be simulated for a region of special interest. If random changes of the rotational parameters within a range smaller than the rotational angle increment are applied to the conformation of the fitted complex, chances are high that near-native conformations are found by the docking procedure. Since the native solution for the benchmarked cases is known, it is possible to increase the number of near-native solutions found by the docking algorithm in this manner. A set of 5,000 predefined rotational parameter combinations, each randomly changed by a maximum of ± 10 degrees, has been used in the way described above. The results for the individual docking runs and the number of near-native conformations respectively are listed in tables 3.1 on the facing page and 3.2 on page 85 for two different cutoff values (4 and 5Å RMSD of the interface C-alpha atoms) along with the changes in accessible surface area (ASA) upon complex

formation for native and fitted unbound complex as calculated with the use of the program NACCESS (Hubbard and Thornton, 1993b).

Table 3.1: Assorted properties and number of near native solutions as found by CKORDO for the test cases of protein-protein docking benchmark 2.0 (Mintseris et al., 2005).

native PDB-ID	ΔASA[a] nat.	ΔASA[b] fit.	$\%ASA$[c] IF	$RMSD$[d] iC_α	# ≤ 4Å[e] 12dg	# ≤ 4Å[e] rRot	# ≤ 5Å[e] 12dg	# ≤ 5Å[e] rRot	SCOP[f] classification		
Rigid-body (63)											
Enzyme-inhibitor / Enzyme-substrate complexes (21)											
1AVX	1585.1	1417.6	8.04	0.47	17	1908	54	1940	b.47.1.2	:	b.42.4.1
1AY7	1237.2	1288.4	12.08	0.54	23	2146	45	2146	d.1.1.2	:	c.9.1.1
1BVN	2221.7	2185.8	9.97	0.87	15	1570	30	1814	b.71.1.1	:	b.5.1.1
1CGI	2052.6	1849.1	12.47	2.02	29	684	103	2544	b.47.1.2	:	g.68.1.1
1D6R	1408.1	1229.4	8.86	1.14	60	256	110	258	b.47.1.2	:	g.3.13.1
1DFJ	2582.0	2619.7	10.28	1.02	2	341	5	843	c.10.1.1	:	d.5.1.1
1E6E	2315.2	1789.7	6.70	1.33	13	2299	24	2541	c.3.1.1	:	d.15.4.1
1EAW	1866.2	1843.5	12.62	0.54	27	318	85	318	b.47.1.2	:	g.8.1.1
1EWY	1501.9	1253.5	6.58	0.8	38	4744	78	5419	b.43.4.2	:	d.15.4.1
1EZU	2751.2	2625.1	10.19	1.21	0	0	0	23	b.16.1.1	:	b.16.1.1
1F34	3038.2	2381.1	11.20	0.93	11	3556	18	3765	b.50.1.2	:	d.62.1.1
1HIA	1736.8	1551.7	10.80	1.4	15	0	86	0	b.47.1.2	:	g.3.15.1
1MAH	2145.5	1876.9	7.85	0.61	14	2459	23	2464	c.69.1.1	:	g.7.1.1
1PPE	1687.8	1731.8	14.70	0.44	92	1503	380	1523	b.47.1.2	:	g.3.2.1
1TMQ	2401.0	2146.2	9.71	0.86	23	4550	37	5358	b.71.1.1	:	a.52.1.2
1UDI	2021.9	2097.7	12.99	0.9	15	615	24	615	c.18.1.1	:	d.17.5.1
2MTA	1461.4	1466.9	6.08	0.41	26	3313	42	3465	b.69.2.1, g.21.1.1	:	b.6.1.1
2PCC	1140.9	1102.2	5.83	0.39	5	715	13	1013	a.93.1.1	:	a.3.1.1
2SIC	1616.8	1603.1	10.11	0.36	28	1814	60	1815	c.41.1.1	:	d.84.1.1
2SNI	1627.9	1348.8	9.42	0.35	28	175	102	175	c.41.1.1	:	d.40.1.1
7CEI	1383.9	1152.8	8.61	0.7	15	1505	30	2460	d.4.1.1	:	a.28.2.1
Antibody-antigen complexes (9)											
1AHW	1899.0	1976.3	6.32	0.69	17	3673	28	4060	b.1.1.1	:	b.1.2.1
1BVK	1321.0	900.5	5.46	1.24	16	429	35	599	b.1.1.1	:	d.2.1.2
1DQJ	1765.0	1502.0	5.88	0.75	4	1031	15	1453	b.1.1.1	:	d.2.1.2
1E6J	1245.5	1090.4	4.49	1.05	38	4361	61	4422	b.1.1.1	:	a.28.3.1
1JPS	1852.3	1923.5	6.44	0.51	18	4717	29	4779	b.1.1.1	:	b.1.2.1
1MLC	1392.0	1225.0	4.80	0.6	6	190	6	190	b.1.1.1	:	d.2.1.2
1VFB	1382.7	1163.2	6.97	1.02	7	605	23	605	b.1.1.1	:	d.2.1.2
1WEJ	1177.5	1069.0	4.24	0.31	7	967	14	967	b.1.1.1	:	a.3.1.1
2VIS	1296.3	869.8	1.33	0.8	0	0	0	0	b.1.1.1	:	b.19.1.2

...continued on next page

Table 3.1 – continued from previous page

native PDB-ID	ΔASA[a] nat.	ΔASA[b] fit.	$\%ASA$[c] IF	$RMSD$[d] iC_α	# ≤ 4Å[e] 12dg	rRot	# ≤ 5Å[e] 12dg	rRot	SCOP[f] classification		
"Other" complexes (22)											
1A2K	1602.7	1429.4	6.27	1.11	9	0	29	0	d.17.4.2	:	c.37.1.8
1AK4	1028.7	1090.9	5.23	1.33	26	513	57	1152	b.62.1.1	:	a.73.1.1
1AKJ	1995.1	1232.2	4.13	1.14	17	2806	31	2855	b.1.1.2	:	b.1.1.1
1B6C	1752.4	1688.2	7.85	1.96	8	2164	15	2445	d.144.1.7	:	d.26.1.1
1BUH	1323.9	1302.9	6.52	0.75	4	1173	11	1335	d.144.1.7	:	d.97.1.1
1E96	1178.8	1179.8	6.04	0.71	7	53	16	81	a.118.8.1	:	c.37.1.8
1F51	2407.2	1680.6	7.48	0.74	24	3078	38	3184	d.123.1.1	:	c.23.1.1
1FC2	1307.1	1085.1	4.10	1.69	8	88	14	88	a.8.1.1	:	b.1.1.2
1FQJ	1806.4	1681.5	7.03	0.91	4	1105	10	1319	a.66.1.1	:	a.91.1.1
1GCQ	1207.7	1063.6	6.31	0.92	16	2098	28	2396	b.34.2.1	:	b.34.2.1
1GHQ	799.9	687.0	3.38	0.34	10	511	23	511	a.102.4.4	:	g.18.1.1
1HE1	2112.8	1773.4	10.82	0.93	8	2095	14	2147	c.37.1.8	:	a.24.11.1
1I4D	1657.2	1445.4	4.64	1.41	0	27	0	84	h.4.7.1	:	c.37.1.8
1KAC	1455.7	1555.1	10.33	0.95	16	2652	30	5076	b.21.1.1	:	b.1.1.1
1KLU	1253.9	1191.2	4.00	0.43	4	8	6	8	b.1.1.2	:	b.40.2.2
1KTZ	989.0	853.8	6.49	0.39	7	432	9	432	g.7.1.3	:	g.17.1.2
1KXP	3341.3	3045.2	7.55	1.12	10	4440	17	5359	a.126.1.1	:	c.55.1.1
1ML0	2069.4	1362.0	3.64	1.02	3	0	9	190	b.116.1.1	:	d.9.1.1
1QA9	1352.5	995.9	4.81	0.73	7	2990	14	3040	b.1.1.1	:	b.1.1.1
1RLB	1438.8	1531.3	5.21	0.66	1	0	10	0	b.3.4.1	:	b.60.1.1
1SBB	1064.1	1225.3	5.08	0.37	26	4255	38	4319	b.40.2.2	:	b.1.1.1
2BTF	2062.5	1708.4	7.21	0.75	0	16	9	363	c.55.1.1	:	d.110.1.1
Antibody-antigen complexes (Crossbound) (11)											
1BJ1	1730.7	1731.7	5.65	0.5	5	2014	5	2068	b.1.1.1	:	g.17.1.1
1FSK	1622.7	1642.1	5.81	0.45	16	4052	25	4052	b.1.1.1	:	d.129.3.1
1I9R	1497.8	1413.8	4.04	1.3	13	4039	21	4205	b.1.1.1	:	b.22.1.1
1IQD	1975.9	1897.0	7.10	0.48	8	222	9	222	b.1.1.1	:	b.18.1.2
1K4C	1600.8	1547.0	4.21	0.53	1	0	1	0	b.1.1.1	:	f.14.1.1
1KXQ	2171.6	2303.9	9.70	0.72	15	2361	31	2476	b.71.1.1	:	b.1.1.1
1NCA	1953.4	1809.7	5.32	0.24	11	4543	16	4545	b.1.1.1	:	b.68.1.1
1NSN	1776.5	1695.7	6.27	0.35	9	2689	17	2967	b.1.1.1	:	b.40.1.1
1QFW^A	1580.5	1382.6	6.05	1.31	10	2381	18	2381	b.1.1.1	:	g.17.1.4
1QFW^B	1636.6	1530.9	7.01	0.73	16	3264	23	3290	b.1.1.1	:	g.17.1.4
2JEL	1500.7	1393.3	5.75	0.17	1	4	5	43	b.1.1.1	:	d.94.1.1
Medium-difficulty (12)											
Enzyme-inhibitor / Enzyme-substrate complexes (2)											
1ACB	1544.0	1735.6	10.83	2.26	12	237	46	265	b.47.1.2	:	d.40.1.1
1KKL	1641.1	1261.4	4.56	2.2	7	920	27	920	c.91.1.2	:	d.94.1.1
Antibody-antigen complexes (Crossbound) (1)											
1BGX	5813.7	5419.1	9.69	1.48	0	0	0	0	a.60.7.1	:	b.1.1.1
"Other" complexes (9)											
1GP2	2286.6	689.6	2.17	1.65	5	0	13	0	b.69.4.1, a.137.3.1	:	a.66.1.1

...continued on next page

Table 3.1 – continued from previous page

native PDB-ID	ΔASA[a] nat.	ΔASA[b] fit.	$\%ASA$[c] IF	$RMSD$[d] iC_α	# ≤ 4Å[e] 12dg	rRot	# ≤ 5Å[e] 12dg	rRot	SCOP[f] classification		
1GRN	2332.2	1570.9	7.88	1.22	19	4331	29	4591	a.116.1.1	:	c.37.1.8
1HE8	1304.9	1086.0	2.39	0.92	3	862	6	1036	a.118.1.6	:	c.37.1.8
1I2M	2779.4	2575.6	9.74	2.12	11	3089	27	4390	b.69.5.1	:	c.37.1.8
1IB1	2807.9	2781.5	8.76	2.09	0	0	0	4	a.118.7.1	:	d.108.1.1
1IJK	1647.9	1322.5	5.87	0.68	8	22	22	32	d.169.1.1	:	c.62.1.1
1K5D	2526.6	1944.7	6.27	1.19	1	83	4	222	c.37.1.8, b.55.1.3	:	c.10.1.2
1M10	2096.6	2246.5	9.85	2.1	0	89	0	129	c.10.2.7	:	c.62.1.1
1WQ1	2913.2	2502.0	9.95	1.16	7	1786	16	2013	a.116.1.2	:	c.37.1.8
Difficult (8)											
"Other" complexes (7)											
1ATN	1774.3	1310.7	4.75	3.28	1	2	3	13	c.55.1.1	:	d.151.1.1
1DE4	2065.5	2032.4	3.07	2.59	1	1199	2	1294	a.48.2.1	:	b.1.1.2
1EER	3346.6	2909.5	8.79	2.44	0	80	5	325	b.1.2.1	:	a.26.1.2
1FAK	3363.1	1833.4	6.66	6.18	0	0	0	0	b.47.1.2, g.3.11.1	:	b.1.2.1
1FQ1	1831.6	1014.6	4.38	3.41	2	1	5	1	d.144.1.7	:	c.45.1.1
1IBR	2070.8	4283.1	13.99	6.62	0	0	0	0	a.118.1.1	:	c.37.1.8
2HMI	3370.4	1206.7	1.79	2.54	0	0	0	0	c.55.3.1, e.8.1.2	:	b.1.1.1
Antibody-Antigen complexes (1)											
1H1V	1234.0	5423.0	10.99	2.26	0	0	0	0	c.55.1.1	:	d.109.1.1

Table 3.2: Assorted properties and number of near native solutions as found by CKORDO for the test cases as collected from the literature.

native PDB-ID	ΔASA[a] nat.	ΔASA[b] fit.	$\%ASA$[c] IF	$RMSD$[d] iC_α	# ≤ 4Å[e] 12dg	rRot	# ≤ 5Å[e] 12dg	rRot	SCOP[f] classification		
Enzyme-Inhibitor/Enzyme-Substrate complexes (21)											
1ACB	14701.04	13157.04	10.5	0.62	42	2749	90	2841	b.47.1.2	:	d.40.1.1
1AVW	17611.63	15871.16	9.88	0.46	31	2001	81	2094	b.47.1.2	:	b.42.4.1
1BRC	12972.12	11655.27	10.15	0.42	34	15	221	15	b.47.1.2	:	g.8.1.1
1BRS	10763.4	9207.46	14.46	0.47	16	1007	48	1007	d.1.1.2	:	c.9.1.1
1BVN	22188.9	19967.18	10.01	0.77	30	1122	88	1122	b.71.1.1	:	b.5.1.1
1CGI	14632.4	12579.83	14.03	1.13	27	940	100	1891	b.47.1.2	:	g.68.1.1
1CHO	13704.27	12237.98	10.7	0.63	53	4266	118	4767	b.47.1.2	:	g.68.1.1
1CSE	13961.98	12473.9	10.66	0.50	87	3350	159	3353	c.41.1.1	:	d.40.1.1
1DFJ	25794.72	23212.72	10.01	1.02	1	0	5	385	c.10.1.1	:	d.5.1.1
1FSS	24100.11	22133.4	8.16	0.64	13	3656	32	3755	c.69.1.1	:	g.7.1.1
1MAH	24273.2	22127.73	8.84	0.67	24	4210	53	4210	c.69.1.1	:	g.7.1.1
1PPF	14162.62	12838.18	9.35	0.47	26	1175	118	2303	b.47.1.2	:	g.68.1.1
1TGS	13489.14	11766.44	12.77	1.19	82	4899	197	4925	b.47.1.2	:	g.68.1.1
1UGH	15375.55	13182.73	14.26	0.52	15	3084	33	4029	c.18.1.1	:	d.17.5.1

...continued on next page

Table 3.2 – continued from previous page

native PDB-ID	ΔASA[a] nat.	ΔASA[b] fit.	$\%ASA$[c] IF	$RMSD$[d] iC_α	# ≤ 4Å[e] 12dg	rRot	# ≤ 5Å[e] 12dg	rRot	SCOP[f] classification		
2KAI	14368.5	12946.66	9.9	0.58	53	1014	263	1014	b.47.1.2	:	g.8.1.1
2MTA	25441.56	23980.17	5.74	0.52	23	3381	50	3400	g.21.1.1, b.69.2.1	:	b.6.1.1
2PCB	19238.35	18208.58	5.35	0.48	2	305	13	1318	a.93.1.1	:	a.3.1.1
2PCC	19259.97	18119.07	5.92	0.42	2	45	13	627	a.93.1.1	:	a.3.1.1
2PTC	13069.55	11640.39	10.94	0.35	99	1709	266	1709	b.47.1.2	:	g.8.1.1
2SIC	15829.68	14212.91	10.21	0.45	36	1261	75	1261	c.41.1.1	:	d.84.1.1
2SNI	14304.33	12676.4	11.38	0.36	18	333	86	333	c.41.1.1	:	d.40.1.1
'Other' complexes (4)											
1AVZ	10020.05	8760.5	12.57	0.55	2	0	8	3	d.102.1.1	:	b.34.2.1
1BDJ	13366.1	12600.77	5.73	0.84	19	726	44	808	c.23.1.1	:	a.24.10.1
1L0Y	22358.78	21226.42	5.06	0.96	11	4040	21	4045	b.1.1.1	:	b.40.2.2
1WQ1	24311.82	21398.67	11.98	0.73	9	568	31	578	a.116.1.2	:	c.37.1.8
Antibody-antigen complexes (4)											
1AHW	29739.47	27772	6.62	0.78	19	3478	30	3823	b.1.1.1	:	b.1.2.1
1VFB	16604.66	15221.99	8.33	0.80	13	1168	33	1168	b.1.1.1	:	d.2.1.2
1WEJ	25715.69	24538.23	4.58	0.63	4	1715	6	1791	b.1.1.1	:	a.3.1.1
1DQJ	25477.66	23712.95	6.93	0.90	1	26	2	29	b.1.1.1	:	d.2.1.2
Difficult (4)											
1BTH	17254.59	14884.37	13.74	0.86	2	0	22	0	b.47.1.2	:	a.74.1.1
1FIN	26390.25	22985.95	12.9	1.26	0	0	0	0	d.144.1.7	:	c.45.1.1
1FQ1	24450.57	22618.93	7.49	1.33	2	1	5	1	d.144.1.7	:	a.137.3.1
1GOT	34520.26	32023	7.23	1.04	0	0	3	18	b.69.4.1, a.137.3.	:	a.66.1.1

Additionally the percentage ratio of the interfacial contact surface as compared to the complete surface of the respective complex is given. Furthermore, the SCOP (Andreeva et al., 2004) (release 1.69 of July 2005) classifications have been retrieved for every distinct chain in the complexes to provide a measure for structural similarity. In tables 3.1 and 3.2 the SCOP distinct classes present in each native complex are listed. These

[a] Change in solvent accessible surface area upon complex formation for the native complex

[b] Change in solvent accessible surface upon complex formation for the conformation of unbound units fitted on the native complex

[c] Percentage ratio of total solvent accessible surface area buried at interface formation

[d] RMSD of interface C-alpha atoms

[e] Number of near native conformations within range of given cutoff ($RMSD$ of interface C_α) found by CKORDO in a) docking run with 12 dg rotational angle sampling (12dg) and b) sampling of 5000 rotations in the range of ±10dg deviation from the conformation of the unbound units as fitted on the native complex (rRot)

[f] SCOP structure classes as present in the complex (classification code(s) in the order of receptor:ligand)

classification codes are identical for the corresponding unbound chains involved.
For the docking benchmark 2.0, the docking algorithm failed to find any near-native conformations in nine cases: Three cases each for the rigid-body, the medium difficulty and the difficult docking categories. For the second, smaller dataset, only one difficult test case did not yield acceptable putative complex orientation below the used cutoff value of 5Å RMSD of interface C_α atoms.
The conformations as listed in the output of CKORDO have subsequently been used to calculate the corresponding values of the employed scoring schemes as listed in section 2.4 via the developed postfilter software. Within the scope of the postfiltering, the number of conformations below a value of 4Å RMSD of interface C-alpha atoms was further increased by the application of slight random movements to already existing conformations within the respectable range of interface C-alpha RMSD. This *artificial enrichment* of near-native conformations was carried out such that an equal number of several thousand near-native solutions for every test case of the two datasets described was reached. This is of critical importance for the next step: the training of probabilistic Support Vector Machines.

3.2 Training and testing of probabilistic Support Vector Machines

3.2.1 Selection of training and testing data

With the existence of two datasets of protein-protein docking test cases as depicted in tables 2.2 and 2.1 along with the respective results from docking experiments (tables 3.1 and 3.2) and the calculated values for the scoring schemes as described in section 2.4, final sets of training and testing data have been selected. Since the *docking benchmark 2.0* represents the far larger, structurally more diverse and non-redundant data, this dataset was used to compose the relevant training data.
The second dataset as collected from various literature resources and manually curated offers docking test cases, which are in large parts structurally similar to those used for training, but not identical to them. Some of the native complexes for the unbound dockings overlap in both datasets, but for the smaller dataset, the unbound units differ

from those used in *docking benchmark 2.0* with the exception of the two Enzyme-Inhibitor test cases for 2SNI and 2SIC. This is the case, since Mintseris et al. (2005) selected the highest quality structure for each of the unbound units, implying that multiple alternatives exist. Additionally, the dockings for the testing data have been conducted starting from randomised starting positions (complex conformations) thus further limiting the chance that any of the interfaces as generated by the docking algorithm are structurally identical to any used as a training instance. The interface structure determines the relative atomic positions as used for the calculation of the scoring schemes which again are used as feature input for training and testing data instances for the supervised machine learning.
The dataset for the *docking benchmark 2.0* is divided into three classes of complexes: Enzyme-Inhibitor and Enzyme-Substrate complexes form the first class, Antibody-Antigen complexes form the second, while every test case not belonging to either of the classes is assigned to the third class depicted as "Other" complexes. This classification has been utilised for the setup of training and testing datasets as well, based on the reasonable assumption that different binding properties distinguish these classes. The functional efficacy of structurally heterogeneous enzymes relies on the specific binding of a rather limited range of substrates/inhibitors while antibodies, a protein class which shares a high structural similarity, bind to a large and diverse set of antigens.
The general strategy for an optimal training of SVMs and sincere testing of their prediction abilities is specified as following:

- Training data
  Constitutes of randomly chosen conformations as taken from docking and post-filter runs of subsets of *docking benchmark 2.0* (cf. table 2.2). Subsets are chosen as defined by the three classes of Enzyme-Inhibitor/Substrate, Antibody-Antigen and "Other" complexes. Data instances included in any of the testing data sets (see next item) are explicitly excluded from the training data.

- Testing data
  Three sets of testing data will be used in order to judge the performance of the trained predictor on unknown data:

  1. Prediction on decoys of *docking benchmark 2.0*:

In order to asses the prediction abilities on those decoys which will presumably be of highest similarity to those used for training, a second set of decoys is randomly chosen from the same pool of conformations as the training data, carefully ensuring that none of the cases used for training is also used for testing.

2. Prediction on decoys of docking examples as collected from various literature resources (cf. table 2.1):
   These test sets are deduced of proposed conformations which have been generated by docking methods of structures that are similar to those used in training. Explicitly, the SCOP classifications as listed in tables 3.1 and 3.2 show a large but no full overlap. Hence this data represents a test set, which is likely to exhibit similarities but no perfect matches to the training data while also including data instances which have been derived from structural data with low or no similarity to the one used for training.
3. Prediction on decoys of docking examples from *docking benchmark 2.0* explicitly excluded from training and previous testing:
   For each of the three docking classes, few complexes have explicitly been excluded from previous training and testing to provide the data for a realistic blind prediction on totally unknown data. Therefore, those complexes have been picked, which share no structural similarity on SCOP superfamily level with any of the complexes used in previous training or testing examples.

Table 3.3: Number of docking test cases and decoys contributing to each of the training and test sets used.

docking class	train		test 1		test 2		test 3	
	# c.[a]	# decoy	# c.	# decoy	# c.	# decoy	# c.	# decoy
Enzyme-Inhibitor/Substrate	23	50,000	23	50,000	21	100,000	2	10,000
Antibody-Antigen	21	50,000	21	50,000	4	20,000	2	10,000
"Other"	29	50,000	29	50,000	4	20,000	2	10,000

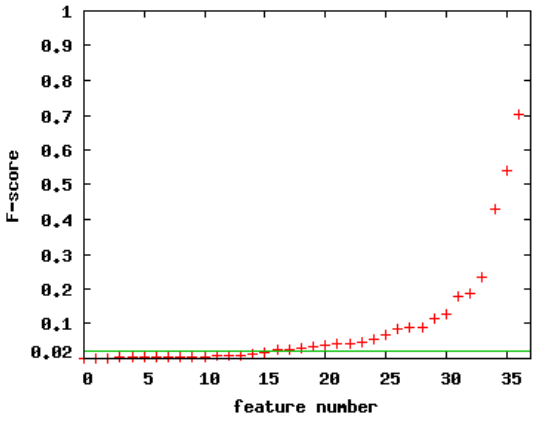
[a]number of complexes for which putative complex conformations as resulting from docking calculations are contributing to the decoy sets

For all the data sets described above, the calculated values for the employed scoring schemes have been normalised using Z-scores (see equation (2.26)). The individual data

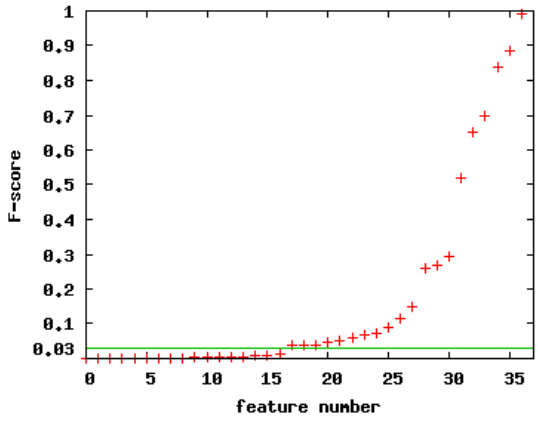
sets for training and testing purposes have been generated by collecting all normalised feature values for each decoy (putative conformation) resulting from the docking runs of the complexes included in the dataset, into a single pool. Data instances within the pool are shuffled and randomly drawn from the ensemble to constitute the training and testing data sets. For all the data sets used in final training and testing, equipartition between near-native and non-near-native (i.e. acceptable and inacceptable) docking solutions was assured. Docking test cases classified as "difficult" test cases were not included into the training or testing data since these proteins are likely to undergo drastic conformational changes upon complex formation which might eventually lead to noisy data for the calculated feature values. Docking test cases for which no near-native conformation could be detected during the docking process have been included into the training data, since also negative or false examples (non-native complex orientations) provide valuable information. The borderline between acceptable, near-native and inacceptable, non-native docking solutions is drawn at 4Å RMSD of interface C-alpha atoms as compared to the unbound units fitted on the native complex.
Table 3.3 gives an overview on the number of docking test cases contributing to each of the training and test sets used.

3.2.2 Feature selection

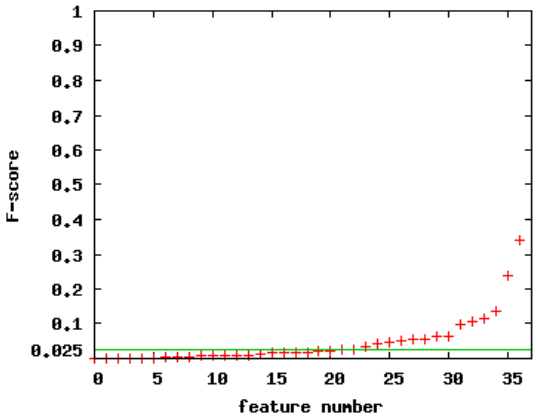
For each of the training data sets as described in the previous section, a feature selection has been performed in order to achieve best possible training results. Therefore, F-scores have been calculated according to equation 2.44 on page 79. Features have been sorted according to the values obtained and the resulting feature numbers (the feature with the highest F-score values was also assigned the highest feature number during sorting) have been plotted against their F-scores (see figure 3.1). On the basis of these plots, cutoff values were manually chosen. Features with F-scores below the cutoff were not included into the training data.
As cutoff for the feature selection according to F-scores, the values 0.02 for the data as originated from Enzyme-Inhibitor/Substrate complexes, 0.03 for the data from Antibody-Antigen complexes and 0.025 for the training set used for "Other" complexes have been selected. This resulted in a total of 21, 20 and 16 features, respectively.
Table 3.4 on page 92 gives an overview of the selected features (highlighted in bold)



(a) Enzyme-Inhibitor/Substrate



(b) Antibody-Antigen



(c) Other

Figure 3.1: F-scores as used for feature selection for the individual training datasets. The green line depicts the manually chosen cutoff value below which features are excluded from training.

along with the corresponding F-scores calculated. The individual scoring schemes are described in this table using the abbreviations as given in the subscript for each scoring value in section 2.4. The probe specific scorings for the GRID-based cost functions are represented by the corresponding directive as used by the GRID software (cf. table 2.3) with a "SE" prefix standing for "solvent effect" (see figure 2.3 and equations (2.11) - (2.12)).

Table 3.4: Feature names and their corresponding F-score values (in order of importance, selected features printed in bold).

	Enzyme-Inh./Substr.		Antib.-Antigen		Other	
#	feat.	F-score	feat.	F-score	feat.	F-score
1	**RPscore**	0.701	**ToF$_{EI}$**	0.991	**RPscore**	0.339
2	**NhcR**	0.540	**RIP$_{AA}$**	0.887	**RIP$_{UNI}$**	0.239
3	**RIP$_{UNI}$**	0.430	**ToF$_{AA}$**	0.837	**COO-**	0.138
4	**BurSurf**	0.232	**RIP$_{EI}$**	0.698	**RIP$_{EI}$**	0.117
5	**RIP$_{EI}$**	0.189	**ToF$_{UNI}$**	0.651	**RIP$_{AA}$**	0.106
6	**GapVol**	0.179	**RPscore**	0.520	**ACE**	0.098
7	**ACE**	0.129	**Cons**	0.295	**N2=**	0.065
8	**RIP$_{AA}$**	0.115	**O-**	0.268	**AMIDINE**	0.062
9	**AMIDINE**	0.088	**ConOE**	0.259	**O::**	0.054
10	**ToF$_{AA}$**	0.087	**RIP$_{EI}$**	0.151	**NhcR**	0.054
11	**N2=**	0.084	**NhcR**	0.113	**N3+**	0.052
12	**ToF$_{UNI}$**	0.068	**NhvR**	0.090	**O-**	0.046
13	**Cons**	0.054	**N3+**	0.074	**ToF$_{UNI}$**	0.042
14	**ToF$_{EI}$**	0.047	**GapVol**	0.068	**O1**	0.034
15	**O-**	0.043	**ACE**	0.060	**CONH2**	0.027
16	**N2**	0.043	**AMIDINE**	0.050	**OH2**	0.025
17	**O**	0.037	**CONH2**	0.045	N1=	0.023
18	**OH2**	0.035	**COO-**	0.038	Cons	0.023
19	**CONH2**	0.031	**N2=**	0.037	BOTH	0.019
20	**NhvR**	0.028	**avgTF**	0.036	OH	0.016
21	**PairPot**	0.026	N1	0.014	SE_BOTH	0.016
22	N1=	0.019	DRY	0.009	avgTF	0.015
23	BOTH	0.014	N1=	0.009	SE_DRY	0.013
24	N1	0.010	O::	0.004	ConOE	0.010
25	SE_DRY	0.009	O	0.004	N2	0.010
26	OH	0.006	N2	0.003	SE_OH2	0.010
27	O::	0.005	SE_DRY	0.003	GapVol	0.009
28	ConOE	0.004	C3	0.002	ToF$_{EI}$	0.007
29	O1	0.003	O1	0.001	PairPot	0.004
30	C1=	0.003	OH2	0.001	DRY	0.004
31	COO-	0.003	OH	0.001	O	0.003
32	C3	0.002	SE_OH2	0.000	NhvR	0.002
33	N3+	0.002	BOTH	0.000	C1=	0.001
34	SE_BOTH	0.002	BurSurf	0.000	N1	0.001
35	SE_OH2	0.002	SE_BOTH	0.000	ToF$_{AA}$	0.001
36	avgTF	0.001	PairPot	0.000	BurSurf	0.000
37	DRY	0.000	C1=	0.000	C3	0.000

3.2.3 Results on training and testing data sets

For each of the training data sets as described in table 3.3 on page 89, Support Vector Machines have been trained using various kernel functions (see equation 2.8.1). For all training data sets, the sigmoidal kernel function provided best possible results. The kernel parameters C and γ have been determined in a grid search procedure using 10-fold cross validation (CV).

Table 3.5 gives an overview of the relevant characteristics and quality measures for the

SVM trainings. The F-score cutoff value, the number of features, the values determined for the kernel parameters C and γ as well as the accuracy reached in the 10-fold cross validation procedure are listed. Additionally, the total number of *Support Vectors* (SV) and the number of *Bounded Support Vectors* (BSV) is given. The higher the percentage of Bounded Support Vectors (given in the rightmost column of table 3.5), the stronger the indication that no overtraining has taken place.

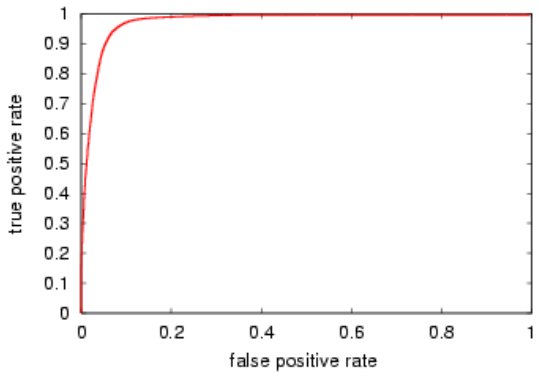
Table 3.5: SVM training characteristics.

docking class	F-score cutoff	# feat.	C	γ	CV-acc.	# SV	# BSV	% BSV
E.-I./S.[a]	0.02	21	1048576	0.000004	93.83	8669	8638	99.64
Ab.-Ag.[b]	0.03	20	32768	0.000977	94.07	8881	8856	99.72
Other[c]	0.025	16	1048576	0.000008	83.02	20532	20507	99.88

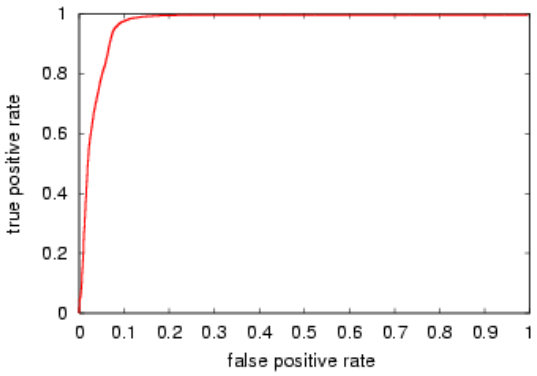
[a]Training data for Enzyme-Inhibitor/Substrate complexes
[b]Training data for Antibody-Antigen complexes
[c]Training data for "Other" complexes

SVM training performance has also been assessed by plotting the rate of true positive against the rate of false negative predicted instances in so called Receiver Operator Characteristic (ROC) plot and calculating the area under the curve. The steeper the curve and the larger the area under the curve (AUC), the better the training performance, respectively. Figure 3.2 shows the ROC plots for the training of SVMs on the three selected complex classes. The corresponding value for the area under the curve are given as well, scaled to a total area of one indicating a perfect predictor, while random guessing would yield an AUC value of 0.5.

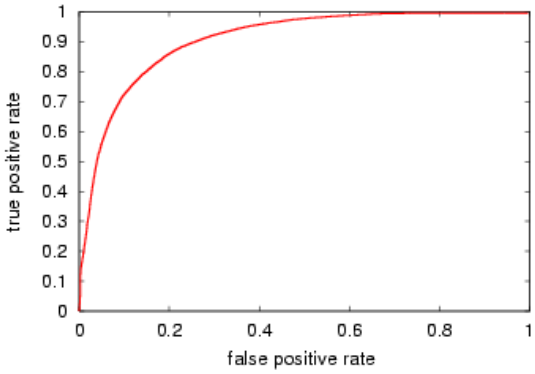
All three trained SVM models have been thoroughly tested on the selected testing datasets as described in section 3.2.1. For each test set, predictions have been performed and the quality measures for machine learning as described in section 2.8.2 on page 76 have been calculated according to equations (2.38) - (2.43). Table 3.6 on page 95 gives an overview of the performance of the trained predictors on the testing datasets. Training and testing quality measures attest a very good performance for the SVM models created for Enzyme-Inhibitor/Substrate and Antibody-Antigen complexes, while the performance for the SVM model trained for the complex class of “Other” complexes is lower relative to the one for the remaining classes.



(a) Enzyme-I./S. complexes, AUC=0.9776



(b) Antib.-Antigen complexes, AUC=0.9703



(c) "Other" complexes, AUC=0.9098

Figure 3.2: Receiver Operating Characteristics (ROC) curves for the training of SVMs on the individual classes of complexes.

These datasets consist of an equal number of true and false solutions each. One has to keep in mind that this is a necessary prerequisite for the significance of some of the relevant performance measures like the Mathews correlation coefficient or the specificity ($spec^-$). On the other hand, this does hardly represent a realistic docking application, where the number of false solutions will always be several orders of magnitude higher than the number of true solutions.

Table 3.6: Quality measures for SVM testing.

Test set	acc	TP	TN	FP	FN	mcc	$spec^+$	$spec^-$	sens	fval
SVM trained on Enzyme-Inhibitor/Substrate complexes (SVM_{EI})										
Test 1	92.43	22450	23764	1236	2550	0.85	0.95	0.95	0.90	0.92
Test 2	83.79	43115	40674	9326	6885	0.68	0.82	0.81	0.86	0.84
Test 3	91.37	4913	4224	776	87	0.84	0.86	0.84	0.98	0.92
SVM trained on Antibody-Antigen complexes (SVM_{AbAg})										
Test 1	90.91	22019	23437	1563	2980	0.82	0.93	0.94	0.88	0.91
Test 2	85.88	9982	7193	2807	18	0.75	0.78	0.72	1.00	0.88
Test 3	91.82	5000	4182	818	0	0.85	0.86	0.84	1.00	0.92
SVM trained on "Other" complexes (SVM_{Oth})										
Test 1	83.14	22067	19501	5499	2933	0.67	0.80	0.78	0.88	0.84
Test 2	51.57	1418	8895	1105	8582	0.05	0.56	0.89	0.14	0.23
Test 3	75.72	4990	2582	2418	10	0.59	0.67	0.52	1.00	0.8

3.3 Scoring protein-protein docking results using probabilistic SVM-classifiers

The three SVM prediction models emerging from the training as described in section 3.2 have been applied to score docked protein complexes. The three predictors have been applied to all the unbound-unbound protein-protein docking test cases from the two benchmarks used in this work with the normalised feature values as calculated by the postfilter software as input. Since probabilistic SVMs are employed, a continuous re-ranking of putative complex conformations yielded by docking calculations is facilitated. Aim of this scoring is the identification of near-native complexes within the top ranks, while inacceptable solutions should emerge on the lower ranks. The borderline for acceptable conformations is drawn at 5Å RMSD of interface C-alpha atoms ($RMSD_{iC_\alpha}$) compared to the unbound units fitted on the native complex. Complexes with an $RMSD_{iC_\alpha}$ below this cutoff values are denoted as "hit". These hits are those complex candidates which can be subjected to a final refinement step. The quality of the SVM-based scoring functions is assessed in the following subsections. A scoring function can only detect a near-native docked complex and subsequently improve the ranking if at least one acceptable solution is present in the decoy set. Consequently, docking test cases for which no near-native solution could be detected were excluded from scoring and evaluation.

3.3.1 Performance of SVM-based scoring functions on designated target complexes

Three SVM-based scoring functions have been developed for specific classes of docking problems each. Consequently, the scoring of docked complexes of the designated target class constitutes the main application focus.

3.3.1.1 Scoring of docked Enzyme-Inhibitor/Substrate complexes

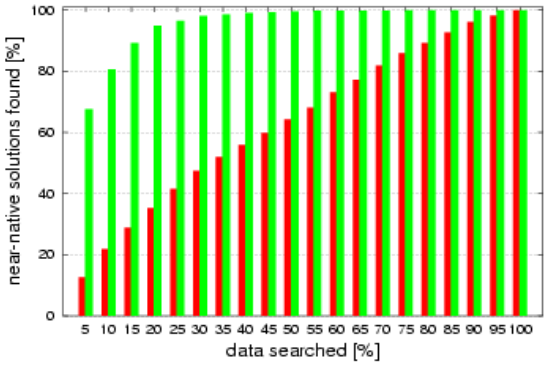
All putative complexes as emerging from docking calculations for the class of Enzyme-Inhibitor/Substrate complexes have been subjected to scoring applying the developed specific SVM-predictor (SVM_{EI}) and subsequent re-ranking. A total of 43 test cases, 22 of the docking benchmark 2.0 (table 2.2) and 21 taken from the manually collected dataset (table 2.1), have been examined. The absolute ranks after scoring according to geometric fit and SVM_{EI} for the complex conformation with the lowest rank (first near-native found) and the complex with the lowest $RMSD_{iC_\alpha}$ value (best near-native found) for the 43 test cases are given in table 3.7. Furthermore, relative performance measures for the SVM-based scoring are listed. This includes the reduction of search space if all putative complexes classified as "false" by the predictor are excluded, the number of false positive solutions per true positive solution, the scoring probability and the improvement factor (see section 2.9 on page 79).

For all 43 test cases, a significant improvement in absolute rank can be noted for both the first near-native complex found as well as the best possible (in terms of $RMSD_{iC_\alpha}$) acceptable solution found. For two of the 43 test cases, a near-native solution could be found within the top ten ranks using the geometric fit as scoring function, while the SVM predictor was able to find an acceptable solution in 21 cases within the top ten ranks. For the top 25 ranks, this ratio totals to five compared to 30 , while in the top 100, 15 compared to 36 cases show a native solution using the geometric and SVM_{EI} scoring, respectively. The average rank could be raised from 827 (1.92% of total ranks) to 84 (0.19%) for the first true and 14,052 (32.62%) to 1,123 (2.61%) for the best true solution comparing geometric to SVM-based ranking.
The mean value for the scoring probability has a rather low value of 0.0639. Compared to the mean scoring probability for the ranking according to geometric fit (0.3820) this

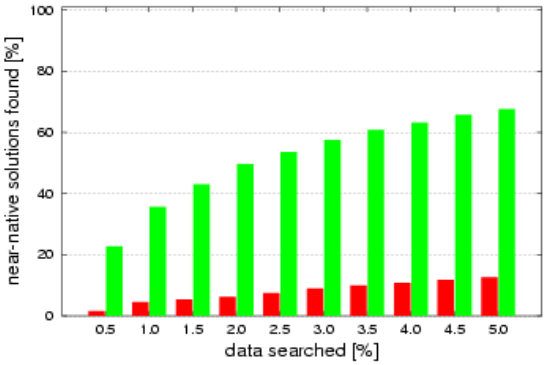
Table 3.7: Performance measures for SVM-based scoring scheme SVM_{EI} on designated target complex class (cyan/magenta: test cases from tables 2.2/2.1).

native PDB-ID	rank 1^{st} near-native				rank best near-native				red.[a]	$\frac{fp}{tp}$[b]	Prob.[c]	IF[d]
	geo[e]	rmsd[f]	pred[g]	rmsd	geo	rmsd	pred	rmsd				
1AVX	1791	4.91	14	4.25	30801	1.52	291	1.52	56.3	355.4	0.0174	2.24
1AY7	1477	4.62	24	3.72	32948	1.39	155	1.39	17.5	789.7	0.0248	1.21
1BVN	61	4.88	2	1.81	370	1.81	2	1.81	71.9	417.5	0.0014	3.43
1CGI	86	4.47	27	4.36	3367	2.74	211	2.74	60.5	165.2	0.0626	2.52
1D6R	591	3.02	768	3.33	41703	1.24	984	1.24	82.4	84.9	0.8621	4.56
1DFJ	3627	3.90	168	3.90	32098	2.77	776	2.77	16.9	7158.0	0.0193	1.20
1E6E	359	3.93	5	2.21	8220	2.01	1503	2.01	36.4	1141.9	0.0028	1.57
1EAW	253	3.48	29	3.39	33867	2.23	3022	2.23	47.3	267.0	0.0557	1.89
1EWY	132	2.15	5	2.15	3662	1.44	44	1.44	70.6	162.4	0.0090	3.38
1F34	51	1.29	1	1.29	51	1.29	1	1.29	49.8	1202.1	0.0004	1.99
1HIA	52	4.43	4	4.51	22393	3.25	12838	3.25	58.8	216.6	0.0079	2.33
1MAH	1175	3.41	3	3.41	8421	1.60	48	1.60	45.1	1029.1	0.0016	1.82
1PPE	151	2.74	1	2.74	10940	1.50	25	1.50	75.6	28.4	0.0088	3.85
1TMQ	19	1.05	3	1.05	19	1.05	3	1.05	24.0	884.9	0.0026	1.31
1UDI	92	3.89	1	3.89	14521	2.03	106	2.03	32.0	1220.5	0.0006	1.47
2MTA	1036	4.92	19	4.92	28712	0.99	317	0.99	20.8	812.2	0.0184	1.26
2PCC	1740	4.17	129	3.90	31193	1.65	497	1.65	16.4	2771.3	0.0382	1.20
2SIC	109	3.01	4	3.89	9457	1.05	22	1.05	38.6	441.1	0.0056	1.62
2SNI	669	4.84	1	3.65	8346	1.95	16	1.95	77.9	101.3	0.0024	4.13
7CEI	60	3.72	13	2.18	6343	1.23	190	1.23	49.8	720.7	0.0090	1.99
1ACB	956	4.57	14	2.79	6349	2.54	532	2.54	79.7	190.5	0.0148	4.89
1KKL	6140	4.81	1250	3.60	16255	1.39	7111	1.39	39.4	1003.6	0.5350	1.65
1ACB	48	0.93	5	0.93	48	0.93	5	0.93	79.4	98.8	0.0104	4.80
1AVW	1440	3.33	6	3.33	24322	1.59	122	1.59	55.7	238.5	0.0297	2.22
1BRC	3	4.42	7	2.56	2236	1.01	13	1.01	63.6	72.3	0.0501	2.66
1BRS	484	4.20	8	2.24	11036	1.97	831	1.97	18.1	766.5	0.0823	1.17
1BVN	32	3.98	9	4.62	3391	1.72	58	1.72	80.2	99.2	0.0282	4.88
1CGI	21	4.54	10	4.81	8506	1.65	75	1.65	67.5	140.0	0.0252	3.05
1CHO	33	1.02	11	1.02	209	0.81	7	0.81	74.4	93.3	0.0055	3.87
1CSE	431	2.26	12	3.53	5107	0.61	124	0.61	61.5	107.6	0.0679	2.50
1DFJ	3634	4.17	192	4.17	18906	3.89	535	3.89	12.1	9463.5	0.0177	1.14
1FSS	2688	1.62	386	4.42	4709	1.09	5037	1.09	26.5	990.0	0.2503	1.36
1MAH	485	4.51	42	2.43	4311	1.18	621	1.18	25.6	604.8	0.0504	1.34
1PPF	164	4.59	5	3.46	23766	2.45	8	2.45	81.4	72.3	0.0136	4.98
1TGS	55	4.21	24	4.53	310	0.85	31	0.85	42.9	125.5	0.1047	1.72
1UGH	127	2.71	3	2.71	31615	2.04	81	2.04	52.2	642.9	0.0022	2.09
2KAI	1	4.63	1	4.63	16772	1.34	1065	1.34	31.4	111.9	0.0061	1.45
2MTA	469	4.74	33	4.27	7905	1.11	165	1.11	22.5	667.9	0.0376	1.29
2PCB	1358	4.49	266	4.43	10399	3.18	8761	3.18	26.3	2644.1	0.0716	1.36
2PCC	1863	4.91	68	4.91	28189	2.88	1096	2.88	10.3	2972.5	0.0203	1.11
2PTC	177	3.81	27	4.23	11774	1.45	764	1.45	9.3	150.3	0.1508	1.10
2SIC	18	3.03	2	3.90	15726	1.08	30	1.08	34.1	417.2	0.0032	1.51
2SNI	1389	4.66	13	4.93	24969	2.87	145	2.87	45.1	353.0	0.0206	1.76
Avg.	827	3.70	84	3.42	14052	1.73	1123	1.73	46.2	976.7	0.0639	2.30

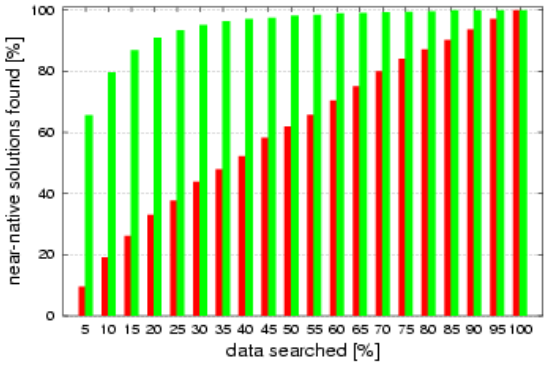
[a]Reduction of data if all conformations classified as "false" are excluded.
[b]Number of flase positive solutions selected per true positive solution.
[c]Scoring probability according to equation (2.46).
[d]Improvement factor according to equation (2.45).
[e]Rank in scoring according to geometric fit.
[f]RMSD of interface C-alpha atoms.
[g]Rank in scoring according to SVM_{EI} predictor.



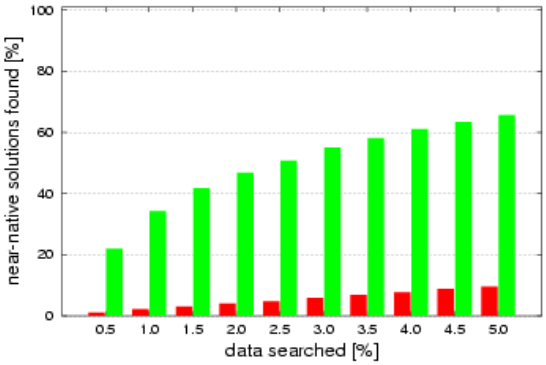
(a) E.-I/S. complexes from table 2.2



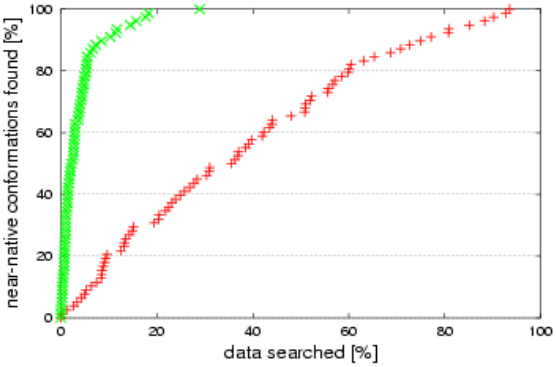
(b) detail of (a): first 5% of ranks



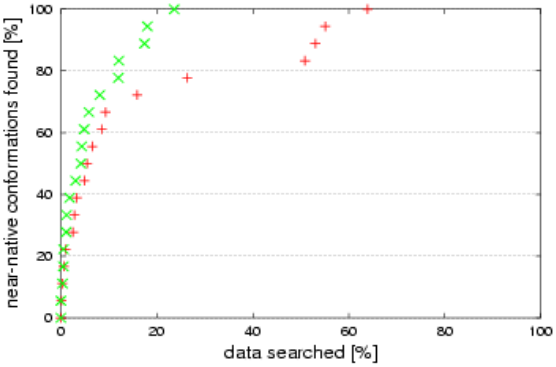
(c) E.-I/S. complexes from table 2.1



(d) detail of (c): first 5% of ranks



(e) blind prediction: test case 1EWY



(f) blind prediction: test case 1F34

Figure 3.3: Enrichment plots depicting the prediction performance of the SVM_{EI} scoring function (green) as compared to the geometric fit (red) on complexes of the target class.

is an improvement by a factor of six. This implies that the chance of obtaining a result as good or better than that obtained by the scoring function by randomly picking complexes out of the pool of generated complexes is six times lower for the SVM-based scoring function than for the scoring according to geometric fit.

For a direct visual comparison of the SVM-based scoring function with the geometric fit, enrichment plots are shown in figure 3.3. Figure 3.3 (a) and (b) illustrate the performance on the complexes of the designated target class of benchmark 2.0. It can clearly be seen that the SVM_{EI} scoring function outperforms the ranking according to geometric fit, accumulating more than 65% of all acceptable solutions within the first 5% of ranks. More than 20% of the near-native solutions can already be detected within the first 0.5% of ranks. The results for the second dataset as collected from the literature (figure 3.3 (c), (d)) are almost identical to those of benchmark 2.0. The performance in blind predictions on test cases of benchmark 2.0 deliberately excluded from training is depicted in figure 3.3 (e), (f). Also here, the developed scoring scheme outperforms the ranking according to geometric fit, with the larger difference in slope of the enrichment curve for example 1EWY, where all acceptable solutions can be found by searching 30% of the data using the SVM-based ranking, while a search of 90% of the data is required if the geometric fit is the primary ranking criterion. For the example of the protein complex 1F34, the total number of acceptable solutions is much smaller and both ranking schemes seem to rank acceptable solutions within the first percentages of the data. The percentage rate of data to be searched in order to find all possible acceptable solutions equals 25% for the SVM_{EI} and 65% for the geometric fit scoring, respectively.

3.3.1.2 Scoring of docked Antibody-Antigen complexes

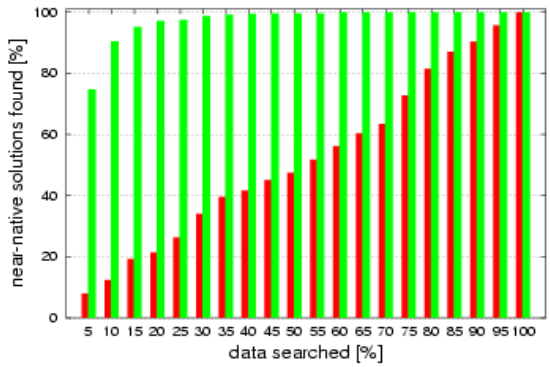
All putative complexes as emerging from docking calculations for the class of Antibody-Antigen complexes have been subjected to a scoring and re-ranking applying the developed specific SVM-predictor for Antibody-Antigen complexes (SVM_{AbAg}). A total of 23 test cases, 19 of the docking benchmark 2.0 (table 2.2) - including 11 cross-bound dockings - and four taken from the manually collected dataset (table 2.1), have been examined. The absolute ranks after scoring according to geometric fit and SVM_{AbAg} for the first near-native complex conformation found and the complex with the lowest

$RMSD_{iC_\alpha}$ value for the 23 test cases are given in table 3.8.

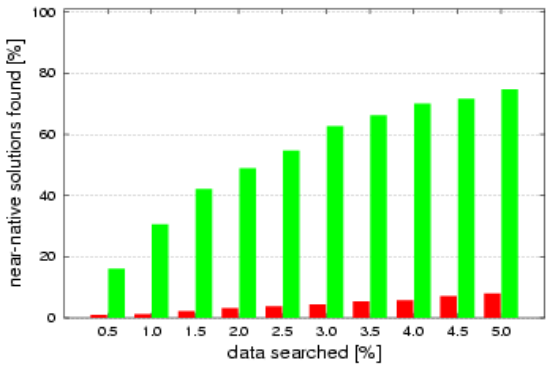
Table 3.8: Performance measures for SVM-based scoring scheme SVM_{AbAg} on designated target complex class (cyan/magenta: test cases from tables 2.2/2.1).

native PDB-ID	rank 1^{st} near-native				rank best near-native				red.	$\frac{fp}{tp}$	Prob.	IF
	geo	rmsd	pred	rmsd	geo	rmsd	pred	rmsd				
1AHW	179	1.89	3	0.91	445	0.91	3	0.91	76.0	382.3	0.0019	4.01
1BVK	1572	4.73	166	3.08	9957	2.23	1808	2.23	41.7	717.0	0.1264	1.71
1DQJ	1965	3.93	606	3.50	39234	3.50	606	3.50	45.0	1578.5	0.1915	1.82
1E6J	453	3.86	46	3.17	16260	1.51	956	1.51	67.9	226.8	0.0631	3.10
1JPS	331	2.01	4	2.01	3362	0.91	12	0.91	81.9	268.6	0.0027	5.51
1MLC	11066	1.93	2053	0.91	20746	0.91	2053	0.91	51.1	3513.8	0.2540	2.04
1VFB	439	4.91	124	4.25	41757	2.50	836	2.50	57.6	795.0	0.0642	2.35
1WEJ	5963	3.77	360	3.51	39767	0.88	489	0.88	75.9	742.8	0.1109	4.14
1BJ1	21571	3.56	244	2.75	24768	2.75	244	2.75	72.4	2373.8	0.0280	3.63
1FSK	1376	2.71	4	1.75	5150	1.75	137	1.75	69.2	530.2	0.0023	3.24
1I9R	468	3.71	15	3.23	2769	1.44	348	1.44	50.0	1024.8	0.0073	2.00
1IQD	4644	3.96	16	3.96	13418	1.22	43	1.22	75.2	1187.8	0.0033	4.03
1K4C	33579	2.43	1081	2.43	33579	2.43	1081	2.43	81.5	7959.0	0.0251	5.41
1KXQ	10	1.90	1306	4.88	1135	1.42	5634	1.42	79.5	440.1	0.6267	3.05
1NCA	1336	3.09	243	2.02	1478	0.73	519	0.73	89.8	293.7	0.0814	9.74
1NSN	815	2.10	12	2.10	23999	1.21	374	1.21	46.1	1221.7	0.0053	1.85
$1QFW^A$	9338	1.60	189	1.60	35118	1.01	839	1.01	72.7	653.7	0.0761	3.66
$1QFW^B$	1893	3.66	95	2.36	14017	0.74	257	0.74	52.3	892.9	0.0495	2.10
2JEL	4922	4.47	87	4.25	5173	3.64	386	3.64	82.4	1517.8	0.0101	5.67
1AHW	20	1.62	1	1.62	2237	1.07	16	1.07	67.1	471.9	0.0007	3.04
1VFB	864	4.89	46	4.76	34295	1.48	679	1.48	68.4	412.3	0.0347	3.16
1WEJ	4306	4.02	213	1.66	21516	0.92	293	0.92	76.2	1707.5	0.0293	4.20
1DQJ	34381	4.39	937	3.98	41524	3.98	937	3.98	35.9	13797.0	0.0430	1.56
Avg.	6152	3.27	341	2.81	18770	1.70	807	1.70	65.9	1856.9	0.0799	3.52

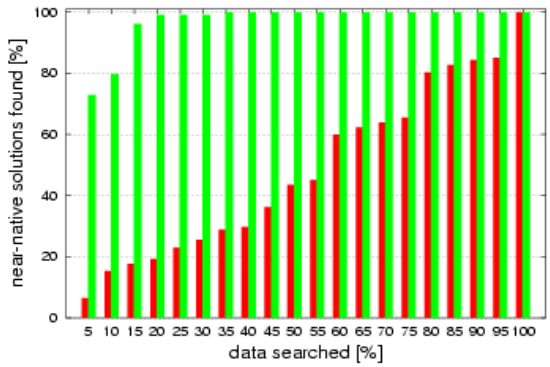
For all but one of the 23 test cases, a significant improvement in absolute rank can be noted for both the first near-native complex found as well as the best possible (in terms of $RMSD_{iC_\alpha}$) acceptable solution. The only exception is the test case 1KXQ of docking benchmark 2.0 for which both criteria experience a significant depletion in absolute rank. For only one of the 23 test cases, a near-native solution could be found within the top ten ranks using the geometric fit as scoring function, while the SVM predictor was able to find an acceptable solution in four cases within the top ten ranks. For the top 25 ranks, this ratio totals to two compared to seven, while in the top 100, two compared to eleven cases show a native solution using the geometric and SVM_{AbAg} scoring, respectively. The average rank could be raised from 6,152 (14.28% of total ranks) to 341 (0.79%) for the first true and 18,770 (43.57%) to 807 (1.87%) for the best true solution comparing geometric to SVM-based ranking.



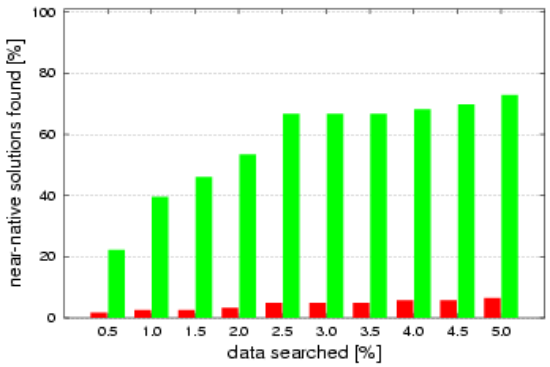
(a) Antib.-Antig. complexes from table 2.2



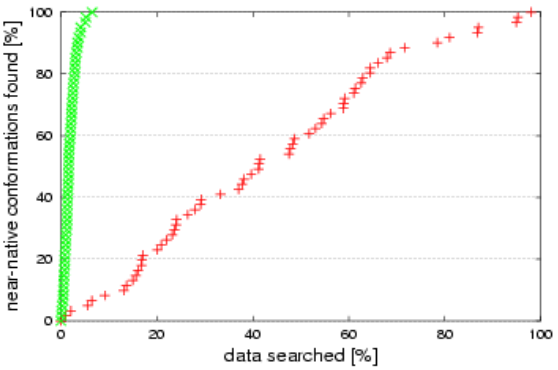
(b) detail of (a): first 5% of ranks



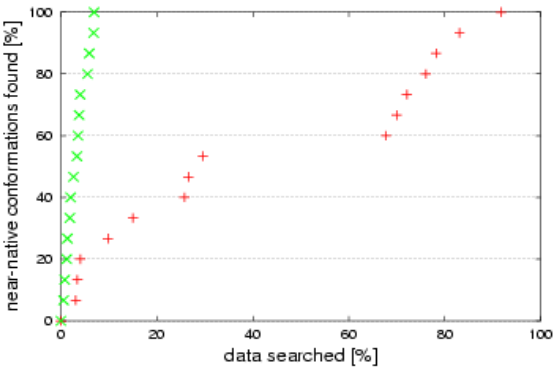
(c) Antib.-Antig. complexes from table 2.1



(d) detail of (c): first 5% of ranks



(e) blind prediction: test case 1E6J



(f) blind prediction: test case 1NCA

Figure 3.4: Enrichment plots illustrating the prediction performance of the SVM_{AbAg} scoring function (green) as compared to the geometric fit (red) on complexes of the target class.

The mean value for the scoring probability of the SVM-based scoring equals 0.0799. Compared to the mean scoring probability for the ranking according to geometric fit (0.5126) this is an improvement by a factor of 6.4. This implies that the chance of obtaining a result as good or better than that obtained by the scoring function by randomly picking complexes out of the pool of generated complexes is more than six times lower for the SVM-based scoring function than for the scoring according to geometric fit.

Compared to the scoring function developed for the class of Enzyme-Inhibitor/Substrate complexes, the mean improvement factor is higher for the SVM_{AbAg}-scoring, though this is mainly due to the fact that the geometric correlation scoring seems to perform worse in the case of Antibody-Antigen complexes.

For a direct visual comparison of the SVM-based scoring function with the geometric fit, enrichment plots are shown in figure 3.4. Figure 3.4 (a) and (b) illustrate the performance on the complexes of the designated target class of benchmark 2.0. The SVM_{AbAg} scoring function clearly outperforms the ranking according to geometric fit, accumulating more than 75% of all acceptable solutions within the first 5% of ranks. More than 15% of the near-native solutions can already be detected within the first 0.5% of ranks. The results for the second dataset as collected from the literature (figure 3.4 (c), (d)) are comparable to those of benchmark 2.0. The performance in blind predictions on test cases of benchmark 2.0 deliberately excluded from training is depicted in figure 3.4 (e), (f). The developed scoring scheme outperforms the ranking according to geometric fit for the test cases 1E6J and 1NCA significantly, ranking all acceptable solutions in the first 10% of the data using the SVM-based ranking, while a search of at least 90% of the data is required if the geometric fit is the primary ranking criterion.

3.3.1.3 Scoring of docked complexes of type "Other" (non-Enzyme-Inhibitor/Substrate and non-Antibody-Antigen complexes)

All putative complexes as emerging from docking calculations for the class of “Other” complexes have been subjected to a scoring and re-ranking applying the developed specific SVM-predictor (SVM_{Oth}). A total of 33 test cases, 29 of the docking benchmark 2.0 (table 2.2) and four taken from the manually collected dataset (table 2.1), have been

examined. The absolute ranks after scoring according to geometric fit and SVM_{Oth} for the first near-native complex conformation found and the complex with the lowest $RMSD_{iC_\alpha}$ value for the 33 test cases are given in table 3.9.

Table 3.9: Performance measures for SVM-based scoring scheme SVM_{Oth} on designated target complex class (cyan/magenta: test cases from tables 2.2/2.1).

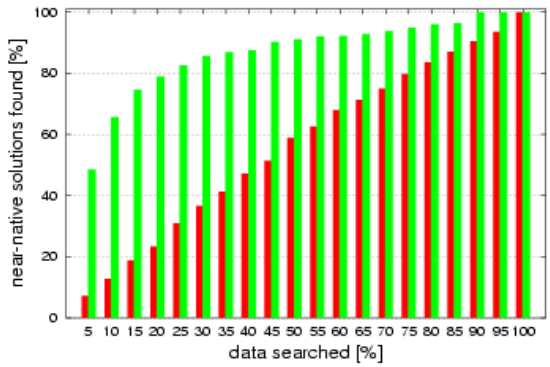
native PDB-ID	rank 1^{st} near-native				rank best near-native				red.	$\frac{fp}{tp}$	Prob.	IF
	geo	rmsd	pred	rmsd	geo	rmsd	pred	rmsd				
1A2K	1463	4.60	139	4.32	10937	3.06	888	3.57	94.6	288.8	0.0895	5.13
1AK4	1787	3.23	29	3.15	9591	2.62	606	2.75	98.2	25.1	0.0377	28.95
1AKJ	733	2.96	167	2.20	3027	2.20	167	2.20	92.1	141.2	0.1135	9.77
1B6C	2464	3.96	1495	3.96	11694	2.36	3603	2.68	85.0	647.4	0.4113	4.43
1BUH	6491	3.83	-	-	15167	2.20	4927	8.19	86.5	-	0.7371	-
1E96	1423	3.91	128	4.69	19389	2.24	865	3.08	78.0	728.5	0.0437	3.94
1F51	1702	1.85	2844	1.99	10180	1.41	6180	1.41	14.8	965.6	0.9255	1.17
1FC2	8020	3.68	3622	3.19	8888	3.19	3622	3.19	79.4	985.9	0.7076	3.12
1FQJ	764	4.69	595	3.59	28966	3.22	733	3.22	72.9	1169.4	0.1299	3.68
1GCQ	2468	2.06	34	4.41	36969	1.33	668	1.33	97.5	74.6	0.0219	20.36
1GHQ	7089	4.83	837	4.90	30490	2.07	6655	2.71	80.2	1066.0	0.3633	1.76
1HE1	691	3.58	35	1.83	6518	1.39	49	1.39	97.0	91.4	0.0113	33.29
1KAC	41	4.17	16	2.96	6735	2.47	74	2.59	92.7	779.2	0.0111	1.84
1KLU	16242	4.66	128	3.37	20504	3.37	128	3.37	61.6	2759.3	0.0177	2.60
1KTZ	14728	3.19	149	2.84	23402	0.82	402	0.82	93.4	317.7	0.0307	15.02
1KXP	14	3.49	11	4.98	54	1.29	29	1.29	77.9	633.9	0.0043	3.99
1ML0	1053	4.47	54	4.47	11339	2.44	848	2.44	25.9	3546.0	0.0112	1.35
1QA9	6412	1.40	71	2.36	6412	1.40	2744	1.40	88.5	352.6	0.0228	8.70
1RLB	5741	4.48	785	3.44	18818	3.44	785	3.44	79.5	981.2	0.1680	4.39
1SBB	876	2.84	-	-	1548	1.22	340	5.70	98.5	-	0.2601	-
2BTF	5849	4.70	168	4.11	19262	3.98	420	3.98	86.8	632.4	0.0346	7.56
1GP2	2018	4.85	238	3.65	35332	1.84	2019	1.84	84.8	502.9	0.0695	6.58
1GRN	195	4.54	551	4.28	9235	1.47	3109	1.47	92.7	195.4	0.3116	7.56
1HE8	1791	4.98	912	4.46	16585	1.67	2780	1.67	88.8	1203.2	0.1205	5.96
1I2M	10	3.80	209	2.41	16582	2.41	209	2.41	84.4	258.5	0.1231	6.15
1IJK	266	4.80	852	4.79	26934	1.70	1062	1.70	60.5	774.1	0.3557	2.53
1K5D	9897	4.97	835	3.35	22980	3.35	835	3.35	36.1	6876.8	0.0753	1.57
1M10	-	-	-	-	37588	5.26	3172	5.26	92.3	-	-	-
1WQ1	909	4.50	286	4.10	21489	1.60	1274	1.60	89.0	338.7	0.1011	7.93
1AVZ	13278	4.51	-	-	21250	3.89	1452	5.74	91.6	-	0.2399	-
1BDJ	186	2.79	703	4.95	6724	1.98	20727	2.08	40.0	957.3	0.5153	1.02
1L0Y	504	2.61	-	-	20301	1.63	635	13.96	98.0	-	0.2680	-
1WQ1	1	3.40	1	2.75	3234	1.71	280	1.71	90.4	138.0	0.0007	10.00
Avg.	3597	3.82	568	3.63	16307	2.31	2191	3.14	79.7	979.6	0.0352	7.51

For 24 of the 33 test cases, a significant improvement in absolute rank can be noted for both the first near-native complex found as well as the best possible acceptable solution. For three test cases (1F51, 1I2M and 1BDJ), the first near-native solution is lowered in rank, while the best near-native solution again is raised in rank. For four test cases (1BUH, 1SBB, 1AVZ, 1L0Y), the SVM predictor fails to classify any

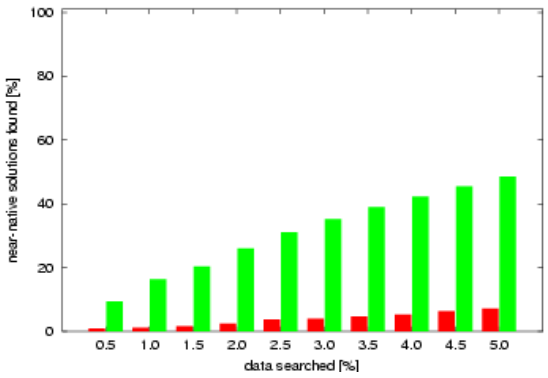
acceptable solution as such. For one further test case in the list (1M10), no near-native solution is found by the docking algorithm and consequently the scoring schemes are destined to "fail". This test is nonetheless listed here in order to show that even if no near-native solution with an RMSD$_i C_\alpha$ of less than 5Å can be found, the SVM-based scoring scheme still provides an improvement in ranking as compared to the geometric fit. For two of the 33 test cases, a near-native solution could be found within the top ten ranks using the geometric fit as scoring function, while the SVM predictor was only able to find an acceptable solution in one case within the top ten ranks. For the top 25 ranks, both scorings succeed in 3 cases, while in the top 100, four compared to eight cases show a native solution using the geometric and SVM$_{Oth}$ scoring, respectively. The average rank could be raised from 3,597 (8.35% of total ranks) to 568 (1.32%) for the first true and 16,307 (37.85%) to 2191 (5.09%) for the best true solution comparing geometric correlation to SVM-based ranking.

The mean value for the scoring probability of the SVM-based scoring equals to 0.0352. Compared to the mean scoring probability for the ranking according to geometric fit (0.5701) this is an improvement by a factor of 16.2. This implies that the chance of obtaining a result as good or better than that obtained by the scoring function by randomly picking complexes out of the pool of generated complexes is more than 16 times lower for the SVM-based scoring function than for the scoring according to geometric fit. Since the SVM$_{Oth}$ scoring scheme fails completely for four of the 33 test cases, this value can only be accepted under reserve, since the high specificity of the scoring is seemingly paid with a lowered sensitivity, explaining the number of test cases for which the scoring scheme fails completely.

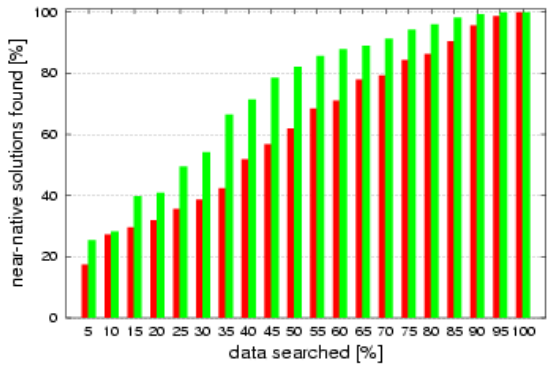
For a direct visual comparison of the SVM-based scoring function with the geometric fit, enrichment plots are shown in figure 3.5 on the facing page. Figure 3.5 (a) and (b) illustrate the performance on the complexes of the designated target class of benchmark 2.0. It can clearly be seen that the SVM$_{Oth}$ scoring function outperforms the ranking according to geometric fit, accumulating 45% of all acceptable solutions within the first 5% of ranks. 10% of the near-native solutions can already be detected within the first 0.5% of ranks. The results for the second dataset as collected from the literature (figure 3.5 (c), (d)) show a significantly worse performance on the four "Other" complexes of table 2.1. Here the discrepancy between SVM-based ranking and geometric correlation is far smaller than the one in plots (a) and (b), though the SVM-based scoring still



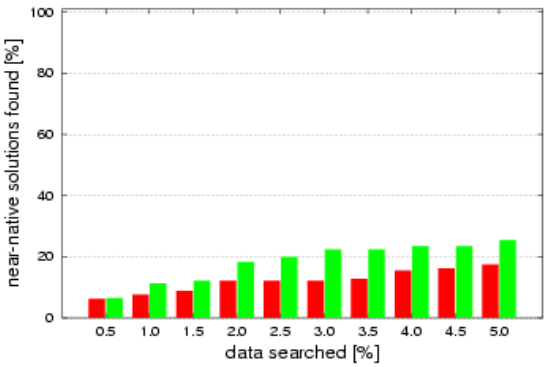
(a) "Other" complexes from table 2.2



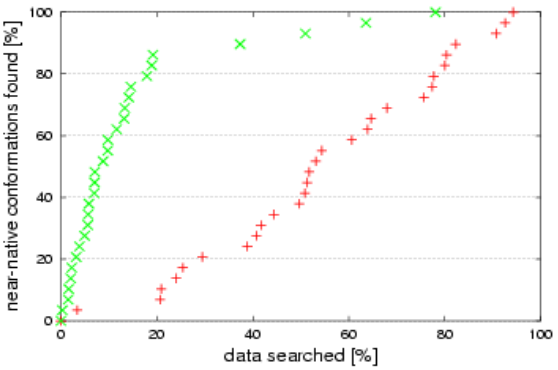
(b) detail of (a): first 5% of ranks



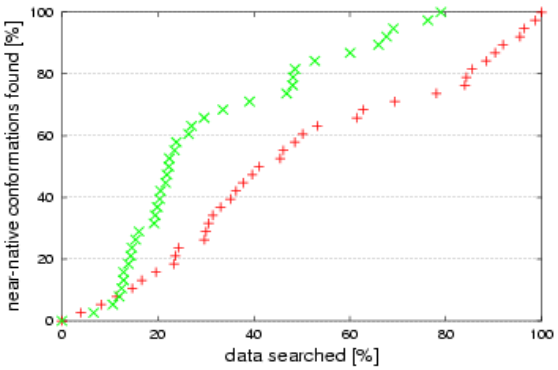
(c) "Other" complexes from table 2.1



(d) detail of (c): first 5% of ranks



(e) blind prediction: test case 1A2K



(f) blind prediction: test case 1F51

Figure 3.5: Enrichment plots illustrating the prediction performance of the SVM_{Oth} scoring function (green) as compared to the geometric fit (red) on complexes of the target class.

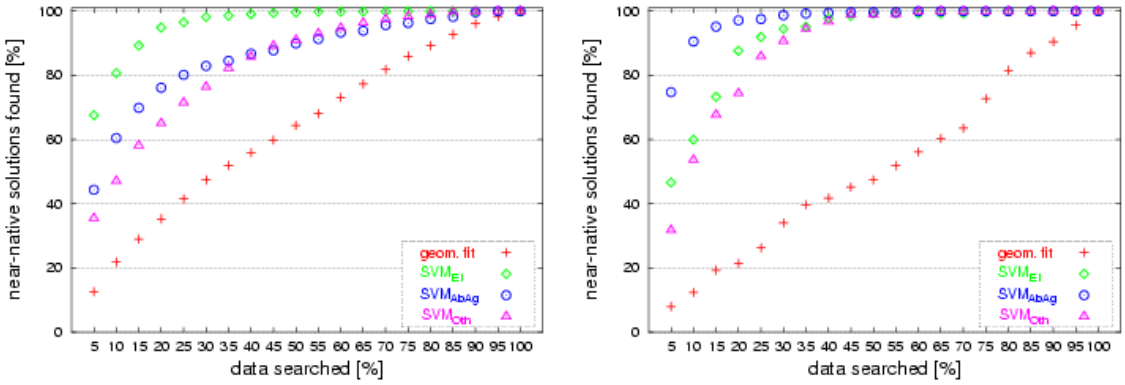
slightly surpasses the geometric fit. The performance in blind predictions on test cases of benchmark 2.0 deliberately excluded from training is depicted in figure 3.5 (e), (f). For test case 1A2K, SVM_{Oth} is working well and clearly superior, while for test case 1F51 it is inferior to the geometric ranking, even if this is only the case for the first 12% of the data.

3.3.2 Specificity of SVM-based scoring functions

The developed SVM-based scoring schemes have been specifically selected and trained for best performance on a designated class of target complexes or rather the docking calculations on these complex classes. In order to assess the specificity of the SVM-based scoring functions for the designated target complex class, each scoring was also applied to those two classes of complexes that were not considered in the training process. Figure 3.6 illustrates a direct comparison of the performance of the specific SVM-based scoring schemes developed on the individual classes of target complexes using enrichment plots.
It can be clearly seen, that for each of the selected complex classes, the SVM-based scoring scheme that has been specifically designed and trained for the appropriate class performs best. Remarkably, all three developed scoring schemes also clearly surpass the scoring according to geometric fit for all complex classes.

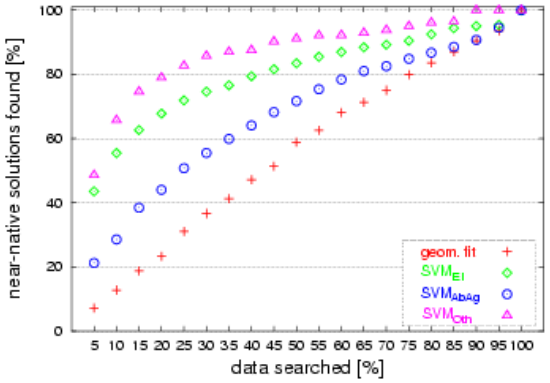
3.3.3 Comparison of SVM-based re-ranking to other scoring functions

So far, the developed scoring schemes have only been compared among each other or to the geometric fit in terms of their performance on the selected classes of complexes or unbound-unbound protein-protein docking test cases, respectively. In order to evaluate the developed SVM-based scoring functions further, a direct comparison of their performance with other common scoring functions was conducted. In total, each complex class specific SVM-based scoring scheme has been compared to five other scoring functions. Namely these are the Atomic Contact Energies (ACE, (Zhang et al., 1997)), a residue-residue potential (RPscore, (Moont et al., 1999)), an atom-atom pair potential (Grimm, 2003) and the class specific scorings using residue interface



(a) Enzyme-Inhibitor/Substrate complexes

(b) Antibody-Antigen complexes

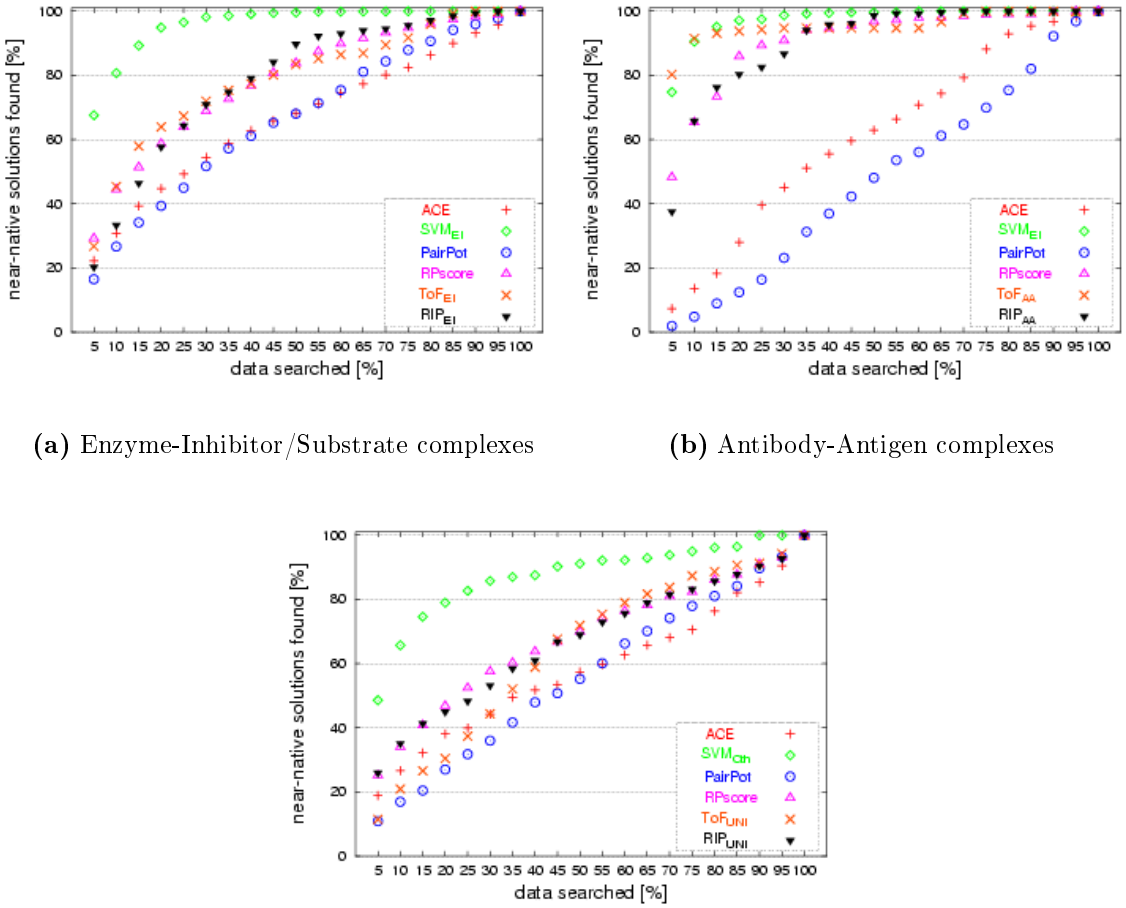


(c) "Other" complexes

Figure 3.6: Enrichment plot illustrating the prediction performance of the three SVM-based scoring functions and the geometric fit for the complexes of benchmark 2.0 (see tables 2.2 and 3.1). The geometric fit is shown as crosses (red), SVM_{EI} as diamonds (green), SVM_{AbAg} as circles (cyan) and SVM_{Oth} as triangles (magenta).

propensities (Huang and Schroeder, 2005) and Tightness of Fit (Gottschalk et al., 2004; Huang and Schroeder, 2005). The results of this comparison of a series of scoring schemes are depicted in figure 3.7 in form of enrichment plots for the three selected classes of docking test cases.

For two of the three classes of docking test cases, the SVM-based scoring scheme features superior performance over all other scoring functions tested. For the classes



(a) Enzyme-Inhibitor/Substrate complexes

(b) Antibody-Antigen complexes

(c) "Other" complexes

Figure 3.7: Enrichment plot illustrating the performance of a series of scoring functions for the complexes of benchmark 2.0 (see tables 2.2 and 3.1). The performance of the Atomic Contact Energies (ACE) scoring is shown as crosses (red), SVM_{EI} as diamonds (green), the atomic pair potential (PairPot) as circles (cyan), the residue potential (RPscore) as triangles (magenta), the Tightness of Fit measure as used together with the residue interface propensities for the specified class of complexes (ToF) in the shape of the letter 'x' (orange), while the mean residue interface propensities for the specified complex class (RIP) are shown as filled inverted triangles (black) .

of Enzyme-Inhibitor/Substrate complexes and "Other" complexes, the discrepancy is quite high. For the class of Antibody-Antigen complexes, all those scoring schemes using residue interface propensities (TOF_{AA}, RPscore, RIP_{AA}) show a general good

performance, with the SVM_{AbAg} and the Tightness of Fit measure surpassing the others. Within the first 10% of ranks, the ToF_{AA} scoring scheme is able to accumulate slightly more near-native solutions than SVM_{AbAg}. In general, the atom based scoring schemes like Atomic Contact Energies (ACE) and the atom-atom pair potential (PairPot) exhibit inferior performance for all three classes of test cases as compared to the other four (residue-level based) scoring functions.

4 Discussion

"It is the mark of an educated mind to be able to entertain a thought without accepting it."

Aristotle, 384-322 B.C.

Efficient comprehensive scoring functions have been developed using probabilistic Support Vector Machines in combination with a series of chemical, biological and physical properties. These scoring functions are shown to be specific for certain types of protein-protein complexes and are able to detect near-native complex conformations from large sets of decoys with high sensitivity. The ranking of near-native structures can be drastically improved, leading to a massive enrichment of near-native complex conformations in the top ranks. It could be shown that the developed scoring schemes outperform five other previously published scoring functions.

4.1 General comparability of docking results

There exists no standardised format for the output, display and evaluation of docking results. Halperin et al. (2002) suggested a unified format for docking results and called it DRUF, the *Docking Results Unified Format*, but this format has so far not become widely accepted in the scientific community. Within the CAPRI challenge, there exists a standardised submission format and a semi-automatic evaluation procedure (Méndez et al., 2003), but these are only usable for and during the actual submission rounds of the challenge.

The obstacles for a standardised docking results format that would allow for a direct comparison of several docking and/or scoring methods arise as early as in the actual setup of the docking test cases. The three dimensional fitting procedure which generates the structural alignment of the unbound units to the native complex is here of critical importance. During the course of this structural alignment, residues of the template have to be assigned to residues in the target structure. This is achieved by a form of structure driven sequence alignment which simultaneously aims for an optimal

matching of residues from the target on those of the template while minimising the overall RMSD. As there is apparently no unique solution to the structural alignment problem (Godzik, 1996), the approaches the methods vary, ranging from fragment based approaches (CE (Shindyalov and Bourne, 1998), PROTEIN3DFIT (Lessel and Schomburg, 1994)) via distance matrices (DALI (Holm and Park, 2000)) to the matching of secondary structure elements (SSM (Krissinel and Henrick, 2004)), just to mention a few. As the methods vary, so do the results in terms of the residue matching between target and template structure (especially in regions where possible gaps have to be introduced), which again influence the calculated overall RMSD. These discrepancies might be small, but for every single step in docking, the RMSD represents the major and most widely used quality criterion, such that eventual discrepancies might add up. Another critical issue is the starting point of the dockings. If the complex conformation used as input structure for the docking procedure corresponds to the structure of the unbound units as fitted on the native complex, a docking algorithm will be more likely to encounter this conformation again during the computation process. Such a near-native solution might be found more easy than if the docking is started from random orientations of the complex partners. Besides the potential discrepancies introduced by methodical variations, many of the definitions used during the quality assessment of docking results are arbitrarily drawn based on empirical knowledge. Examples for this are the setting of borderlines that divide acceptable near-native solutions from inacceptable, non-native ones. Is a docking solution with an RMSD of 3Å, 4Å or 5Å indicated as near native? Is the RMSD calculated with respective to the native complex or rather to the unbound units as fitted on the native complex? Are all atoms used for the calculation of the RMSD or only backbone atoms, eventually only C-alpha atoms? Is the RMSD calculated for the complete protein or rather only the interface region? If so, how is the *interface region* defined? Another frequently used quality criterion is the number of native and non-native contacts on residue or atom level, where again one has to ask "how these are defined". In order to compare various scoring or re-ranking methods for protein-protein docking solutions, it is not only important that the parameters as described above are in agreement but also that they are applied to a possibly identical set of decoys. Only if the number of putative complex conformations produced by a docking algorithm is identical for the docking targets considered, absolute rank numbers, as allocated sometimes, can be directly

compared. This problem has been addressed in this work by the utilisation of relative scoring performance criteria such as enrichment plots, improvement factors and scoring probabilities.

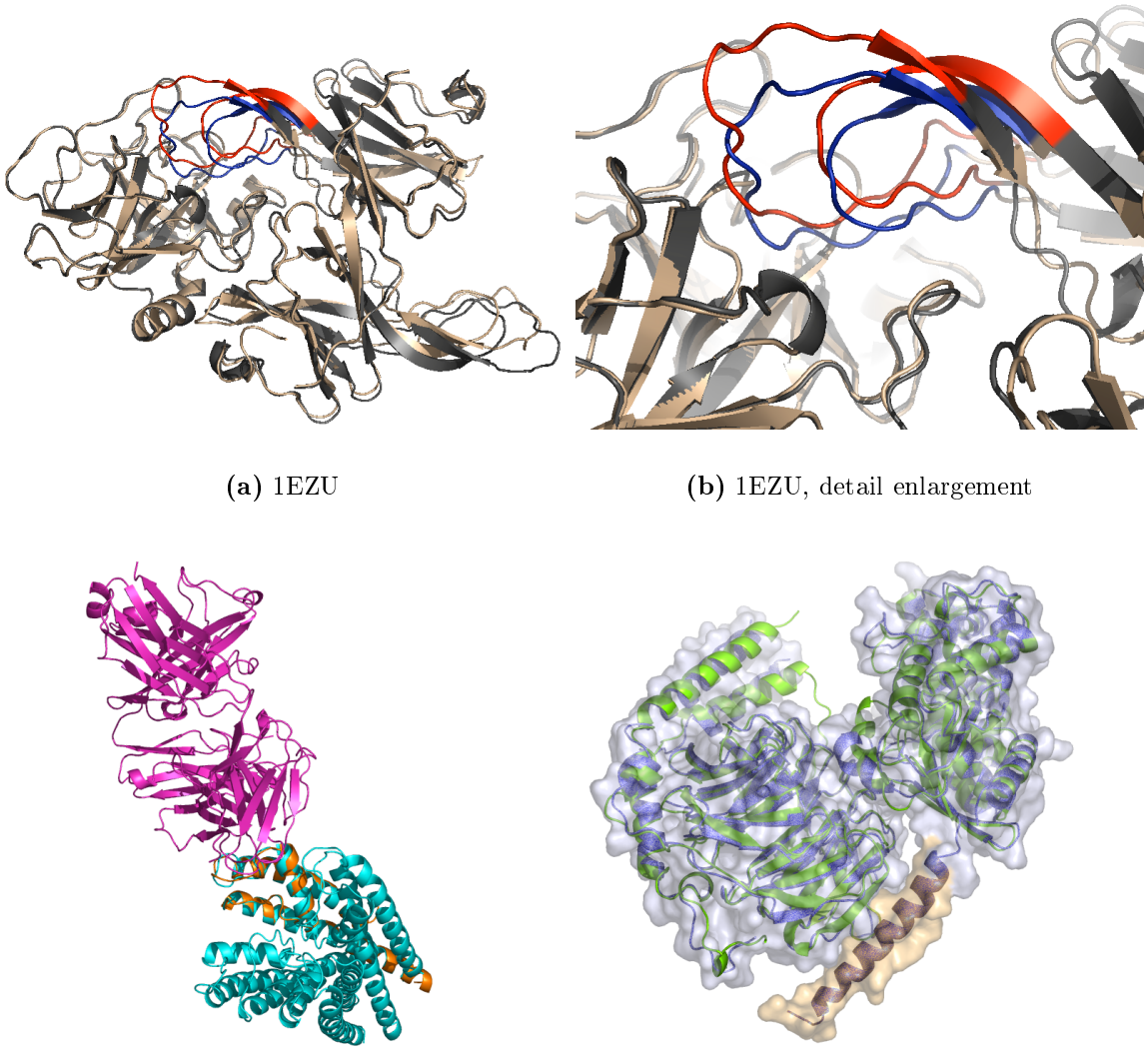
4.2 Limitations of data fundamentals and docking software

For this work, the largest currently available dataset for protein-protein docking has been employed. Still the number of complexes used in the studies comes to 118, 83 of which can be considered as structurally non-redundant interactions. This represents only a very limited fraction of the currently available structural data. These 83 complexes are constituted of 101 unique SCOP domains, originating from 82 folds of eight distinct folding classes, with 94 superfamilies and 101 families represented. Compared to SCOP version 1.69, which is based on a hierarchical clustering of 25,973 PDB entries, this only covers 8.68% of the folds, 5.90% of the superfamilies and 3.55% of the families of non-redundant structural data available at the time (October 2004). Consequently, there can be no guarantee that any knowledge derived from studies of such limited data fundamental as the docking research community relies on can be successfully transferred to future examples.

Another major issue is of course not only the quantity of the data but also its quality as seen in combination with the weak spots of the docking methods/software used. While reconstructing the data as listed in the protein-protein docking *benchmark 2.0*, namely generating the conformation of the unbound units as fitted on the native complex using the structures as deposited in the Protein Data Bank, it became obvious that the authors had to manually curate at least some of the files in order to transform them into a suitable docking test case.

For seven test cases, multimers were created from the PDB files, four of which (1EZU, 1K4C, 1IB1, 1BGX) lead to docking problems with C_n symmetry axes involving multiple symmetric solutions. Currently, few of the available docking programs (e.g. M-ZDOCK (Pierce et al., 2005) and SYMMDOCK (Schneidman-Duhovny et al., 2005b)) are able to handle such symmetric multimer dockings. Unfortunately, CKORDO is in the existing version not able to cope with such problems and will therefore only be

able to identify the $\frac{1}{n}$th part of the near native solutions of a C_n symmetric multimer docking test case.



(a) 1EZU

(b) 1EZU, detail enlargement

(c) 1K4C

(d) 1GP2

Figure 4.1: Problematic docking test cases from benchmark 2.0.

The dataset further contains test cases for which the interface area as seen in relation to the total surface of the complex is relatively small. For 20 of the 83 examined test cases the contact surface covers less than 5% of the total surface of the complex, whereas the

value drops below 3% for four test cases. CKORDO seems to give poor results in terms of the absolute number of near-native solutions listed for these docking runs. Since the docking software used represents plain rigid-body docking, all those test cases which either undergo drastic conformational changes during complex formation or which exhibit a large number of steric clashes for the fitted complex pose a problem. Examples for conformational changes are hinge movements (as in 1FAK), loop movements in the interface area (1EZU) or also conformational changes on an implicit level, caused by atoms which contribute to the interaction in the native complex but are missing in the respective unbound units (1GP2). An aggregation of steric clashes (for the fitted unbound units) in the interface region can be noted with the test cases for 1H1V and 1BGX.

Figure 4.1 on the page before illustrates the problems described on selected examples: Figure 4.1 (a) and (b) show the docking test case for 1EZU from the docking benchmark 2.0, a docking test case labeled as *rigid-body*. The native complex (gray) and the fitted unbound units (brown) differ critically on the receptor side of the interface. A pair of loops exhibits a bad fit or loop movement respectively between the native (highlighted in blue) and unbound state (highlighted in red) as can be clearly seen in the detail enlargement (figure 4.1 (b)). Figure 4.1 (c) shows the docking test case for 1K4C from the docking benchmark 2.0, a multimeric docking test case. For the co-crystallised complex (ligand: orange, receptor: magenta), a single interface is visible, while for the unbound units, a multimer has been generated in order to simulate the biological unit. The unbound receptor (cyan) exhibits a C_4 symmetry, providing for four identical docking site or four equally correct docking solutions, respectively. Of these four solutions, CKORDO is only able to recognize one, since the reference complex for RMSD calculations only exhibits a single "native" interface and symmetries cannot be considered in the current version. In figure 4.1 (d), the medium difficulty docking test case for 1GP2 is depicted with the native complex in blue (cartoon and surface view) and the fitted unbound units in green (cartoon representation only). The native receptor exhibits an additional long alpha helix (surface highlighted in orange) which contributes to the interface. This helix is missing in the unbound receptor, presumably since this part of the structure is likely to be either highly flexible or most definitely no to be found in an orientation similar to the one of the bound state.

4.3 Quality of the developed comprehensive scoring functions

It has been the aim of this work to develop methods for the efficient and accurate detection of near-native conformations in the scoring or ranking process of docked protein-protein complexes. For best possible performance, the scoring function(s) have been developed specifically for distinct classes of protein-protein complexes based on the assumption, that distinct binding preferences distinguish the corresponding complex classes.

4.3.1 Effects of feature selection on the specificity of the SVM-based scoring functions

The results of the performance comparison of the SVM-based scoring functions on the distinct classes of protein complexes (see section 3.3.2) clearly show, that each developed scoring function performs best in the ranking of docked complexes of the class it has been specifically developed for. This substantiates the hypothesis that the determining properties for binding differ between the complex classes in their relevancy and magnitude. Certain conclusions on those binding properties might be drawn from the F-score values as calculated during the feature selection process (cf. table 3.4).
Comparing the F-score values for the individual complex classes, it is apparent that the number of highly conserved residues is seemingly of higher importance for the class of Enzyme-Inhibitor/Substrate than for Antibody-Antigen complexes. This is in agreement with the findings of Reddy and Kaznessis (2005) that the surface density of highly conserved positions is significantly higher in interface regions of protein-protein complexes which do not belong to the class of Antibody-Antigen complexes. This is to be expected since the variable region of the antibody represents the interacting region with the antigen and indeed, the number of highly variably residues reaches its largest F-score value for the class of Antibody-Antigen complexes.
Furthermore, the residue interface propensity scores using propensity values as determined for specific classes of complexes correlate well in their relevance for the prediction of binding sites according to the calculated F-score values with the classes of complexes they have been calculated for. The residue interface propensity score using propensities

deducted from Enzyme-Inhibitor complexes (RIP_{EI}) is of larger importance for the class of Enzyme-Inhibitor/Substrate than for the other two complex classes, while the residue interface propensity score for Antibody-Antigen complexes (RIP_{AA}) dominates in the feature selection of Antibody-Antigen complexes and the universal residue interface propensity score (RIP_{UNI}), using interface propensities deducted from a heterogeneous data set, works best for the most heterogeneous class of docking test cases labeled as "Other" complexes.

The atom-atom pair potential only plays an important role for the training of Enzyme-Inhibitor/Substrate complexes. This is explicable since the potential has been developed using the COMBASE (Vakser and Sali, 1999) dataset, in which the class of Enzyme-Inhibitor complexes is clearly overrepresented.

The buried surface area and the gap volume are used by the existing version of CKORDO to calculate the gap index as a measure for the tightness of binding. It has been tested mainly on Enzyme-Inhibitor complexes so far and also in this work, the scores for gap volume and buried surface area have the biggest influence in terms of F-score values for the class of Enzyme-Inhibitor/Substrate complexes.

Regarding the GRID based scoring schemes, it is apparent that for all three docking complex classes examined, a *minimum set* of probes has been selected. Such a minimum set contains at least one positively charged group (AMINDINE, N3+) and one negatively charged group (COO-, O-), one hydrogen donor (N2=, N2, CONH2) and one hydrogen acceptor group (CONH2, O, O::, O1). A minimum set of probes is consequently able to account for the most important electrostatic and hydrogen bonding forces. Additionally, the solvent probes for water (0H2) and a hydrophilic solvent (DRY) are usually part of a minimum set of probes. The solvent probe for water is only part of the selected features for the classes of Enzyme-Inhibitor/Substrate and "Other" complexes, but is not included for the class of Antibody-Antigen complexes. Interestingly, for none of the complex classes, a hydrophobic probe (DRY, C1=, C3) has been selected due to the low F-score values. This means that the created predictors do not account for purely hydrophobic interaction forces directly and might pose a potential weakness of the trained SVM-based scoring functions.

4.3.2 Effects of training data selection on the quality of the SVM-based scoring functions

Since protein-protein docking calculations typically yield a very low number of acceptable solutions among a large number of false solutions, it is of utmost importance that any developed scoring function misses to identify as few near-native conformations as possible. This means that not only the specificity of a scoring function should be a decisive quality criterion but also its sensitivity. While the SVM-based scoring functions trained for the scoring of docked Enzyme-Inhibitor/Substrate and Antibody-Antigen complexes show almost perfect sensitivity in training and testing (see sections 3.2.3 and 3.3.1), the sensitivity of the SVM-based scoring function for the class of "Other" complexes clearly has deficits concerning the sensitivity. The most likely reason for this lack of sensitivity is the heterogeneity of the complex structures grouped as "Other" complexes and used as input data. This becomes already obvious during the feature selection and SVM training procedure. The average F-score value of the 37 features used lies 60% below the mean values reached for the other two classes, indicating a far weaker descriptive power for the classification of the features for the class of "Other" complexes. While approximately 15% of the data instances are transformed into Support Vectors for the classes of Enzyme-Inhibitor/Substrate and Antibody-Antigen complexes, the total number of Support Vectors reaches $> 40\%$ of the training data instances in the case of "Other" complexes without indication of an eventual overtraining (cf. table 3.5). An eventual improvement of the sensitivity of the SVM_{Oth} scoring function could be achieved by increasing the number of training instances, which again is limited by the low data fundamentals.

4.4 Support Vector Machines as black box

As with every machine learning technique, one has to be aware, that the algorithm will learn to distinguish the given datasets on the basis of the provided descriptors only. In this case, this implies that the SVM procedure has learned to distinguish near-native from inacceptable docking solutions on the basis of the scoring schemes used as descriptors. It is possible but not essential that the machine learning has thereby implicitly learned about phenomena that govern the principles of protein-

protein interaction. Considering the limitations of the raw data as discussed above, this is a legitimate concern. Since there is no way of judging from an SVM model (i.e. the number and sizes of the support vectors) on what actually happens in the high dimensional feature space, Support Vector Machines act as the proverbial *black box*.

Another important point to keep in mind is the relatively simple feature selection method that had to be chosen due to the high computational effort of the SVM training caused by the kernel function utilised and the number of parameters and training instances. Feature selection via F-scores is a fast method but bears the disadvantage that it does not reveal mutual information among features. This mutual information can only be revealed if SVMs are actually trained on a combination of the respective features, like it would have been possible by the application e.g. of a genetic algorithm for feature selection.

4.5 Versatility of the developed method

During the development of the method described in this work, special attention has been paid to the the versatility with respect to future applications. Therefore a bias by CKORDO and its parameters was possibly avoided. The developed postfiltering method deliberately passes on any direct shape complementarity scores or other scorings as calculated by the FFT methods. It should therefore be independent of the underlying method of conformational space search and insensitive to eventual changes to the latter.

5 Conclusion and outlook

It has been the aim of this work to develop methods for the efficient and accurate detection of near-native conformations in the scoring or ranking process of docked protein-protein complexes. A series of structural, chemical, biological and physical properties are employed to score docked protein-protein complexes. These scoring schemes include specialised probe specific energy functions, evolutionary relationship, class specific residue interface propensities, the gap volume, buried surface area, empiric pair potentials on residue and atom level as well as measures for the tightness of fit. Using the largest currently available benchmark of protein-protein docking test cases, supervised machine learning algorithms in the form of probabilistic Support Vector Machines have been trained after feature selection to establish efficient comprehensive scoring functions specific for three different classes of protein-protein complexes. These docking classes are Enzyme-Inhibitor/Substrate complexes, Antibody-Antigen complexes and a third class covering all those complexes not belonging to either of the two previous classes. The three specific scoring functions were tested on the docking results of 43, 23 and 33 complexes in their unbound form for the above mentioned complex classes and are shown to be specific for the individual types of complexes. Defining success as scoring a 'true' result with a p value of better than 0.1, the scoring schemes were found to be successful in 93%, 78% and 63% of the examined cases, respectively. A comparison with five previously published scoring schemes showed the developed class specific comprehensive scoring functions to be superior to the individual scoring functions and illustrated the synergetic effect.

5.1 Future developments

In the era of structural genomic initiatives which expedite the worldwide effort of automised high throughput structure elucidation, the number of known protein sequences is still growing much faster than the number of known structures. Therefore, vast interest is focused on methods which are able to predict protein structures with high accuracy on one hand as well as algorithms for interaction prediction such as

docking programs which facilitate reliable results using modelled structures on the other hand. Another center of attention is focused on data integration trying to relate information about proteins from heterogeneous experimental and theoretical resources aiming for a final complete detailed description of protein interaction networks.
As recent developments of the CAPRI docking prediction challenge show, where first dockings with modelled structures are currently assessed, docking has come of age. It has evolved from a purely academic experiment, being solely able to reliably predict bound docking test cases, to a promising field of practical (research) applications.
The methodology developed in this work is in its current version applicable as a protein-protein docking post filter. Since special attention has been paid to the versatility of the method (see 4.5), one further aim should be the integration of the primary docking software and the developed post filter together with other developments into a fully functional docking software suite. In order to be practically applicable to a wide range of docking problems, such a software should allow for the integration of various external data resources. Examples are:

- on macroscopic level:
  Informations concerning complex symmetries (e.g. from sequence or structural homologies, Electron Microscopy or low resolution X-ray experiments etc.) leading to multimer docking algorithms,

- on microscopic level:
  Integration of information on interaction restraints (e.g. information identifying interface residues (e.g. from cross-linking, mutagenesis, NMR or Mass Spectrometry experiments etc.) or non-interface residues.

The currently biggest challenge in docking software development is the integration of flexibility into the docking methods in order to account for potentially large structural changes upon complex formation. At the same time, methods should still be accurate on atomic level while being tolerant against structural deviations as emerging from modelled structures. This is only possible if sophisticated simulation and refinement methods are combined and integrated into a docking application which preferably should be able to dock, evaluate and refine whole ensembles of structures in a reasonable amount of time. Utilisation of recent threading and parallelisation techniques can facilitate such costly computations.

References

Albrecht, C., Blank, K., Lalic-Mülthaler, M., Hirler, S., Mai, T., Gilbert, I., Schiffmann, S., Bayer, T., Clausen-Schaumann, H., and Gaub, H. E. (2003). DNA: a programmable force sensor. *Science*, 301(5631):367–70. 1.3.1.3

Alpaydin, E. (2004). *Introduction to Machine Learning*. The MIT Press. 2.8

Althaus, E., Kohlbacher, O., Lenhof, H.-P., and Müller, P. (2002). A combinatorial approach to protein docking with flexible side chains. *J Comput Biol*, 9(4):597–612. 1.4.2.4

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10. 2.1

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402. 2.4.7

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32(Database issue):D226–9. 2.1, 2.4.2, 3.1

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(96):223–30. 1.3.2.2

Arya, S., Mount, D., Netanyahu, N., Silverman, R., and Wu, A. (1998). An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM*, 45:891–923. 2.4.1, 2.7

Ausiello, G., Cesareni, G., and Helmer-Citterich, M. (1997). Escher: a new docking procedure applied to the reconstruction of protein tertiary structure. *Proteins*, 28(4):556–67. 1.4.2.1

Aytuna, A. S., Gursoy, A., and Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21(12):2850–5. 1.4.2.3

Back, J. W., de Jong, L., Muijsers, A. O., and de Koster, C. G. (2003). Chemical cross-linking and mass spectrometry for protein structural modeling. *J Mol Biol*, 331(2):303–13. 1.3.1.1

Bahadur, R. P., Chakrabarti, P., Rodier, F., and Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, 336(4):943–55. 1.2.1.4

Baldi, P. and Brunak, S. (1998). *Bioinformatics The Machine Learning Approach*. The MIT Press. 2.8

Ben-Zeev, E. and Eisenstein, M. (2003). Weighted geometric docking: incorporating external information in the rotation-translation scan. *Proteins*, 52(1):24–7. 1.4.2.3

Ben-Zeev, E., Kowalsman, N., Ben-Shimon, A., Segal, D., Atarot, T., Noivirt, O., Shay, T., and Eisenstein, M. (2005). Docking to single-domain and multiple-domain proteins: old and new challenges. *Proteins*, 60(2):195–201. 1.4.2.2

Bennett, M. J., Choe, S., and Eisenberg, D. (1994). Domain swapping: entangling alliances between proteins. *Proc Natl Acad Sci U S A*, 91(8):3127–31. 1.1.2, 1.1.3

Berchanski, A., Shapira, B., and Eisenstein, M. (2004). Hydrophobic complementarity in protein-protein docking. *Proteins*, 56(1):130–42. 1.4.2.3

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28(1):235–42. 1.1.1

Betts, M. J. and Sternberg, M. J. (1999). An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng*, 12(4):271–83. 1.4.2.3, 1.4.2.4

Bhaskar, H., Hoyle, D. C., and Singh, S. (2005). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Comput Biol Med.* 2.8

Bock, J. R. and Gough, D. A. (2001). Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–60. 1.3.2.2

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31(1):365–70. 2.4.7

Bogan, A. A. and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol*, 280(1):1–9. 1.2, 1.2.1.2

Bonvin, A. M. (2006). Flexible protein-protein docking. *Curr Opin Struct Biol.* 1.4.3

Bonvin, A. M., Boelens, R., and Kaptein, R. (2005). NMR analysis of protein interactions. *Curr Opin Chem Biol.* 1.3.1.2

Boobbyer, D. N., Goodford, P. J., McWhinnie, P. M., and Wade, R. C. (1989). New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J Med Chem*, 32(5):1083–94. 2.3

Boser, B., I. G. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory.* 2.8.1

Bradford, J. R. and Westhead, D. R. (2003). Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem. *Protein Sci*, 12(9):2099–103. 1.3.2.2

Bradford, J. R. and Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–94. 1.3.2.3

Brooks, B., Bruccoleri, R. E., Olafson, B. D., States, D. J., SSwiminathan, S., and Karplus, M. (1983). *J Comp Chem*, 4:187. 2.3

Burges, C. J. C. (2002). *2002*. Kluwer Academic Publishers, Boston. 2.5.1.3

Camacho, C. J. and Vajda, S. (2001). Protein docking along smooth association pathways. *Proc Natl Acad Sci U S A*, 98(19):10636–41. a

Carter, P., Lesk, V. I., Islam, S. A., and Sternberg, M. J. E. (2005). Protein-protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI. *Proteins*, 60(2):281–8. 1.4.2.4

Carugo, O. and Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Sci*, 6(10):2261–3. 1.2.1.4

Chakrabarti, P. and Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–43. 1.2.1.1, 1.2.1.2, 1.2.1.3, 2.4.2

Chang, C.-C. and Lin, C.-J. (2005). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/" "cjlin/libsvm`. 2.8.1

Chen, R., Li, L., and Weng, Z. (2003a). ZDOCK: an initial-stage protein-docking algorithm. *Proteins*, 52(1):80–7. 1.4.2.2

Chen, R., Mintseris, J., Janin, J., and Weng, Z. (2003b). A protein-protein docking benchmark. *Proteins*, 52(1):88–91. 2.1, j

Chen, R. and Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, 47(3):281–94. c, 2.4.6

Chen, Y.-W. and Lin, C.-J. (2004). Combining SVMs with various feature selection strategies. In Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., editors, *Feature extraction, foundations and applications*. Springer. 2.8.3

Clausen-Schaumann, H., Rief, M., Tolksdorf, C., and Gaub, H. E. (2000). Mechanical stability of single DNA molecules. *Biophys J*, 78(4):1997–2007. 1.3.1.3

Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–13. 1.4.2.1

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–8. 1.3.2.1

Date, S. V. and Marcotte, E. M. (2005). Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatics*, 21(10):2558–2559. 1.3.2.1

Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125(7):1731–7. 1.4.3

Duan, Y., Reddy, B. V. B., and Kaznessis, Y. N. (2005). Physicochemical and residue conservation calculations to improve the ranking of protein-protein docking solutions. *Protein Sci*, 14(2):316–28. 1.4.2.3

Duncan, B. S. and Olson, A. J. (1993). Shape analysis of molecular surfaces. *Biopolymers*, 33(2):231–8. 1.4.2.1

Eisenstein, M. and Katchalski-Katzir, E. (2004). On proteins, grids, correlations, and docking. *C R Biol*, 327(5):409–20. 1

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90. 1.3.2.1

Fancy, D. A. (2000). Elucidation of protein-protein interactions using chemical cross-linking or label transfer techniques. *Curr Opin Chem Biol*, 4(1):28–33. 1.3.1.1

Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002). Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem*, 269(5):1356–61. 1.3.2.3

Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2004). Identification of protein-protein interaction sites from docking energy landscapes. *J Mol Biol*, 335(3):843–65. 2.6

Fernández, A. and Scheraga, H. A. (2003). Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc Natl Acad Sci U S A*, 100(1):113–8. 1.4.2.3

Fernández-Recio, J., Totrov, M., and Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins*, 52(1):113–7. 1.4.3

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–6. 1.3.1.1

Fischer, E. (1894). Einuss der konguration auf die wirkung der enzyme. *Chem.Ber.*, 27:2985–2993. 1.4.3

Fitzjohn, P. W. and Bates, P. A. (2003). Guided docking: first step to locate potential binding sites. *Proteins*, 52(1):28–32. 1.4.3

Gabb, H. A., Jackson, R. M., and Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*, 272(1):106–20. 1.4.2.2, 1.4.2.3, d

Gabdoulline, R. R. and Wade, R. C. (1998). Brownian dynamics simulation of protein-protein diffusional encounter. *Methods*, 14(3):329–41. 1.4.2.3

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

Gardiner, E. J., Willett, P., and Artymiuk, P. J. (2003). GAPDOCK: a Genetic Algorithm Approach to Protein Docking in CAPRI round 1. *Proteins*, 52(1):10–4. 1.4.2.3

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–7. 1.3.1.1

Gillilan, R. E. and Lilien, R. H. (2004). Optimization and dynamics of protein-protein complexes using B-splines. *J Comput Chem*, 25(13):1630–46. 1.4.2.4

Glaser, F., Pupko, T., Paz, I., Bell, R., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19(1):163–4. 2.4.7

Godzik, A. (1996). The structural alignment between two proteins: is there a unique answer? *Protein Sci*, 5(7):1325–1338. 4.1

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem*, 28(7):849–57. 2.3, 2.4.1, 2.4.1

Gottschalk, K.-E., Neuvirth, H., and Schreiber, G. (2004). A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *Protein Eng Des Sel*, 17(2):183–9. 1.4.2.3, 2.4.4, 2.4.8, 2.9, 3.3.3

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*, 331(1):281–99. 1.4.3

Grimm, V. (2003). *Untersuchung eines wissensbasierten Potentials zur Bewertung von Protein-Protein-Docking-Studien*. PhD thesis, Universität zu Köln. 1.4.2.3, 2.4.5, 3.3.3

Grünberg, R., Leckner, J., and Nilges, M. (2004). Complementarity of structure ensembles in protein-protein binding. *Structure (Camb)*, 12(12):2125–36. 1.4.2.4

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–43. 1.4.2, 1.4.3, i, 2.6, 4.1

Halperin, I., Wolfson, H., and Nussinov, R. (2004). Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure (Camb)*, 12(6):1027–38. 1.2.1.2, 1.4.2.3

Heifetz, A., Katchalski-Katzir, E., and Eisenstein, M. (2002). Electrostatics in protein-protein docking. *Protein Sci*, 11(3):571–87. f

Heuser, P., Baù, D., Benkert, P., and Schomburg, D. (2005). Refinement of unbound protein docking studies using biological knowledge. *Proteins*. 1.4.2.3

Holm, L. and Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics*, 16(6):566–567. 4.1

Honig, B. and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–9. 1.4.2.3

Hopfinger, A. J. (1973). *Conformational Properties of Macromolecules*, page Chapter2. Academic Press, New York, USA. 2.3

Hu, Z., Ma, B., Wolfson, H., and Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins*, 39(4):331–42. 1.2.1.2

Huang, B. and Schroeder, M. (2005). Using residue propensities and tightness of fit to improve rigid-body protein-protein docking. In Matthias Rarey, Andrew Torda, S. K. and Willhoeft, U., editors, *Proceedings of German Bioinformatics Conference*. Springer. 1.4.2.2, 1.4.2.3, 2.4.2, 2.5, 2.9, 3.3.3

Hubbard, S. and Thornton, J. (1993a). Naccess, computer program. *Department of Biochemistry and Molecular Biology, University College London*. 2.4.9

Hubbard, S. and Thornton, J. (1993b). *'NACCESS' Computer Program*. Department of Biochemistry and Molecular Biology, University College London. 3.1

Jackson, R. M., Gabb, H. A., and Sternberg, M. J. (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J Mol Biol*, 276(1):265–85. 1.4.2.4

Jackson, R. M. and Sternberg, M. J. (1995). A continuum model for protein-protein interactions: application to the docking problem. *J Mol Biol*, 250(2):258–75. 1.4.2.3

Janin, J. (2005). Assessing predictions of protein-protein interaction: the CAPRI experiment. *Protein Sci*, 14(2):278–83. 1.4.4

Janin, J. and Chothia, C. (1990). The structure of protein-protein recognition sites. *J Biol Chem*, 265(27):16027–30. 1.2.1.1

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I., and Wodak, S. J. (2003). Capri: a critical assessment of predicted interactions. *Proteins*, 52(1):2–9. 1.4.4, 2.6

Janin, J. and Rodier, F. (1995). Protein-protein interaction at crystal contacts. *Proteins*, 23(4):580–7. 1.2.1.4

Jiang, F. and Kim, S. H. (1991). "soft docking": matching of molecular surface cubes. *J Mol Biol*, 219(1):79–102. 1.4.2.1, 1.4.2.2

Jones, S., Marin, A., and Thornton, J. M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng*, 13(2):77–82. 2.4.2

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20. 1.2

Jones, S. and Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol*, 272(1):121–32. 1.2.1.1, 1.2.1.3, 2.4.2

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637. 2.2.1

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A*, 89(6):2195–9. 1.4.2.1, 1.4.2.2, 2.2.1, 2.4.1

Kawahashi, Y., Doi, N., Takashima, H., Tsuda, C., Oishi, Y., Oyama, R., Yonezawa, M., Miyamoto-Sato, E., and Yanagawa, H. (2003). In vitro protein microarrays for detecting protein-protein interactions: application of a new method for fluorescence labeling of proteins. *Proteomics*, 3(7):1236–43. 1.3.1.1

Koch, C. A., Anderson, D., Moran, M. F., Ellis, C., and Pawson, T. (1991). SH2 and SH3 domains: elements that control interactions of cytoplasmic signaling proteins. *Science*, 252(5006):668–74. 1.3.2.2

Koch, K., Zöllner, F., Neumann, S., Kummert, F., and Sagerer, G. (2002). Comparing bound and unbound protein structures using energy calculation and rotamer statistics. *In Silico Biol*, 2(3):351–68. 1.4.2.4

Krissinel, E. and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 12 Pt 1):2256–2268. 4.1

Krissinel, E. B., Winn, M. D., Ballard, C. C., Ashton, A. W., Patel, P., Potterton, E. A., McNicholas, S. J., Cowtan, K. D., and Emsley, P. (2004). The new CCP4 Coordinate Library as a toolkit for the design of coordinate-related applications in protein crystallography. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 12 Pt 1):2250–5. 2.7

Krämer, P. (2001). *Ermittlung, Charakterisierung und effziente Verarbeitung von Oberflächenparametern für die Simulation von molekularen Wechselwirkungen der Proteine*. PhD thesis, Universität zu Köln. 1.4.2.3

Landschulz, W. H., Johnson, P. F., and McKnight, S. L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science*, 240(4860):1759–64. 1.3.2.2

Lanman, J. and Prevelige, P. E. (2004). High-sensitivity mass spectrometry for imaging subunit interactions: hydrogen/deuterium exchange. *Curr Opin Struct Biol*, 14(2):181–8. 1.3.1.1

Larsen, T. A., Olson, A. J., and Goodsell, D. S. (1998). Morphology of protein-protein interfaces. *Structure*, 6(4):421–7. 1.2

Laskowski, R. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph*, 13(5):323–30, 307–8. 2.2.1, 2.4.10

Lessel, U. and Schomburg, D. (1994). Similarities between protein 3-D structures. *Protein Eng*, 7(10):1175–1187. 4.1

Li, C. H., Ma, X. H., Chen, W. Z., and Wang, C. X. (2003a). A soft docking algorithm for predicting the structure of antibody-antigen complexes. *Proteins*, 52(1):47–50. 1.4.3

Li, L., Chen, R., and Weng, Z. (2003b). RDOCK: refinement of rigid-body protein docking predictions. *Proteins*, 53(3):693–707. 1.4.2.4

Lichtarge, O., Bourne, H., and Cohen, F. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2):342–58. 1.3.2.2

Lichtarge, O. and Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol*, 12(1):21–7. 1.3.2.2

Lijnzaad, P. and Argos, P. (1997). Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins*, 28(3):333–43. 1.2.1.2

Lijnzaad, P., Berendsen, H. J., and Argos, P. (1996). A method for detecting hydrophobic patches on protein surfaces. *Proteins*, 26(2):192–203. 1.2.1.2

Lin, S. L., Nussinov, R., Fischer, D., and Wolfson, H. J. (1994). Molecular surface representations by sparse critical points. *Proteins*, 18(1):94–101. 1.4.2.1

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–98. 1.2.1.2, 1.2.1.3, 1.4.2.2, 2.4.2, 2.5

Lorber, D. M., Udo, M. K., and Shoichet, B. K. (2002). Protein-protein docking with multiple residue conformations and residue substitutions. *Protein Sci*, 11(6):1393–408. g

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A*, 100(10):5772–7. 1.2.1.2

Ma, B., Wolfson, H. J., and Nussinov, R. (2001). Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Curr Opin Struct Biol*, 11(3):364–9. 1.2.1.2

Ma, X. H., Li, C. H., Shen, L. Z., Gong, X. Q., Chen, W. Z., and Wang, C. X. (2005). Biologically enhanced sampling geometric docking and backbone flexibility treatment with multiconformational superposition. *Proteins*, 60(2):319–23. 1.4.3

Malmqvist, M. (1993). Biospecific interaction analysis using biosensor technology. *Nature*, 361(6408):186–7. 1.3.1.3

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., and Eyck, L. F. T. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Eng*, 14(2):105–13. 1.4.2.2, 1.4.2.3, h

Mangasarian, O. L. (2001). Data mining via support vector machines. Technical report, University of Wisconcin, Computer Sciences Department. 2.5.1.3

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–3. 1.3.2.1

Masters, S. C. (2004). Co-immunoprecipitation from transfected cells. *Methods Mol Biol*, 261:337–50. 1.3.1.1

McCoy, A. J., Epa, V. C., and Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces. *J Mol Biol*, 268(2):570–84. 1.4.2.3

Melo, F. and Feytmans, E. (1997). Novel knowledge-based mean force potential at atomic level. *J Mol Biol*, 267(1):207–22. (document), 1.4.2.3, 2.4.1, 2.1, 2.4

Meyer, M., Wilson, P., and Schomburg, D. (1996). Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. *J Mol Biol*, 264(1):199–210. 1.4.2.3, 2.2.1

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., and Weng, Z. (2005). Protein-Protein Docking Benchmark 2.0: an update. *Proteins*, 60(2):214–6. (document), 2.1, 2.2, 3.1, 3.2.1

Miyazawa, S. and Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, 18:534–552. 2.4.6

Miyazawa, S. and Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256(3):623–44. 1.4.2.3

Moont, G., Gabb, H. A., and Sternberg, M. J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, 35(3):364–73. 1.4.2.3, 2.4.3, 3.3.3

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*, 15(3):285–9. 1.4.4

Méndez, R., Leplae, R., Lensink, M. F., and Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60(2):150–69. 1.4.4

Méndez, R., Leplae, R., Maria, L. D., and Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, 52(1):51–67. 1.4.4, 2.1, 4.1

Neuvirth, H., Raz, R., and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*, 338(1):181–99. 1.3.2.3, 2.4.8

Neves-Petersen, M. T. and Petersen, S. B. (2003). Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules–applications in biotechnology. *Biotechnol Annu Rev*, 9:315–95. 1.4.2.3

Nooren, I. M. A. and Thornton, J. M. (2003a). Diversity of protein-protein interactions. *EMBO J*, 22(14):3486–92. 1.1.1, 1.1.1

Nooren, I. M. A. and Thornton, J. M. (2003b). Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol*, 325(5):991–1018. 1.1.1, 1.1.3

Norel, R., Fischer, D., Wolfson, H. J., and Nussinov, R. (1994). Molecular surface recognition by a computer vision-based technique. *Protein Eng*, 7(1):39–46. 1.4.2.2

Ofran, Y. and Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS Lett*, 544(1-3):236–9. 1.3.2.2

Palma, P. N., Krippahl, L., Wampler, J. E., and Moura, J. J. (2000). Bigger: a new (soft) docking algorithm for predicting protein interactions. *Proteins*, 39(4):372–84. e

Park, J., Lappe, M., and Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol*, 307(3):929–38. 2.4.2

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 271(4):511–23. 1.3.2.2

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–14. 1.3.2.1

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. R., CheathamIII, T. E., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995). Amber, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp. Phys. Commun.*, 91:1–41. 2.2.1

Pearlman, D. A. and Charifson, P. S. (2001). Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *J Med Chem*, 44(21):3417–23. 1.4.2.3

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–8. 1.3.2.1

Pierce, B., Tong, W., and Weng, Z. (2005). M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics*, 21(8):1472–8. 4.2

Pierce, M. M., Raman, C. S., and Nall, B. T. (1999). Isothermal titration calorimetry of protein-protein interactions. *Methods*, 19(2):213–21. 1.3.1.3

Platt, J. (2000). *Advances in Large Margin Classifiers*, chapter Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. MIT Press. 2.8.1.1

Pupko, T., Bell, R., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 Suppl 1:S71–7. 2.4.7

Rausa, F. M., Hughes, D. E., and Costa, R. H. (2004). Stability of the hepatocyte nuclear factor 6 transcription factor requires acetylation by the CREB-binding protein coactivator. *J Biol Chem*, 279(41):43070–6. 1.1.2

Reddy, B. V. B. and Kaznessis, Y. N. (2005). A quantitative analysis of interfacial amino acid conservation in protein-protein hetero complexes. *J Bioinform Comput Biol*, 3(5):1137–50. 2.4.7, 4.3.1

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–2. 1.3.1.1

Ritchie, D. W. and Kemp, G. J. (2000). Protein docking using spherical polar fourier correlations. *Proteins*, 39(2):178–94. 1.4.2.1

Rittinger, K., Walker, P. A., Eccleston, J. F., Nurmahomed, K., Owen, D., Laue, E., Gamblin, S. J., and Smerdon, S. J. (1997a). Crystal structure of a small G protein in complex with the GTPase-activating protein rhoGAP. *Nature*, 388(6643):693–7. 1.1.1

Rittinger, K., Walker, P. A., Eccleston, J. F., Smerdon, S. J., and Gamblin, S. J. (1997b). Structure at 1.65 A of RhoA and its GTPase-activating protein in complex with a transition-state analogue. *Nature*, 389(6652):758–62. 1.1.1

Rodgers, G. P. (1997). Overview of pathophysiology and rationale for treatment of sickle cell anemia. *Semin Hematol*, 34(3 Suppl 3):2–7. 1.1.2

Scarsi, M., Majeux, N., and Caflisch, A. (1999). Hydrophobicity at the surface of proteins. *Proteins*, 37(4):565–75. 1.4.2.3

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005a). Geometry-based flexible and symmetric protein docking. *Proteins*, 60(2):224–31. 1.4.3

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005b). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res*, 33(Web Server issue):W363–7. 4.2

Schneidman-Duhovny, D., Inbar, Y., Polak, V., Shatsky, M., Halperin, I., Benyamini, H., Barzilai, A., Dror, O., Haspel, N., Nussinov, R., and Wolfson, H. J. (2003). Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins*, 52(1):107–12. 1.4.3

Schölkopf, B. (1997). *Support Vector Learning*. R. Oldenbourg Verlag, Munich. 2.8.1.1

Schölkopf, B., Guyon, I., and Weston, J. (2003). *Artificial Intelligence and Heuristic Methods in Bioinformatics*, chapter 1, pages pages 1–21. IOS Press. Statistical learning and kernel methods in bioinformatics. 2.5.1.3

Sear, R. P. (2004). Highly specific protein-protein interactions, evolution and negative design. *Phys Biol*, 1(3-4):166–172. 1.1.2

Sheinerman, F. B. and Honig, B. (2002). On the role of electrostatic interactions in the design of protein-protein interfaces. *J Mol Biol*, 318(1):161–77. 1.4.2.3

Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng*, 11(9):739–47. 4.1

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213(4):859–83. 1.4.2.3

Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K., and Zhu, G. (2003). Improved amino acid flexibility parameters. *Protein Sci*, 12(5):1060–72. 2.4.8

Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–7. 1.3.1.1

Smith, G. R., Fitzjohn, P. W., Page, C. S., and Bates, P. A. (2005a). Incorporation of flexibility into rigid-body docking: applications in rounds 3-5 of CAPRI. *Proteins*, 60(2):263–8. 1.4.2.4, 1.4.3

Smith, G. R., Sternberg, M. J. E., and Bates, P. A. (2005b). The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J Mol Biol*, 347(5):1077–101. 1.4.2.4

Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, C. (2000). *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA. 2.8.1.1

Sprinzak, E., Sattath, S., and Margalit, H. (2003). How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327(5):919–23. 1.3.1.3

Susnow, R. G. and Dixon, S. L. (2003). Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J Chem Inf Comput Sci*, 43(4):1308–1315. 2.5.1.1

Taylor, J. S. and Burnett, R. M. (2000). Darwin: a program for docking flexible molecules. *Proteins*, 41(2):173–91. 1.4.3

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80. 2.4.7

Tress, M., de Juan, D., Graña, O., Gómez, M. J., Gómez-Puertas, P., González, J. M., López, G., and Valencia, A. (2005). Scoring docking models with evolutionary information. *Proteins*, 60(2):275–80. 1.4.2.3

Vajda, S. and Camacho, C. J. (2004). Protein-protein docking: is the glass half-full or half-empty? *Trends Biotechnol*, 22(3):110–6. 1.4.4

Vakser, I. and Sali, A. (1999). http://salilab.org/sub-pages/combase.html. Unpublished data. 2.4.5, 4.3.1

Vakser, I. A. (1995). Protein docking for low-resolution structures. *Protein Eng*, 8(4):371–7. 1.4.3

Vakser, I. A. (1996). Low-resolution docking: prediction of complexes for underdetermined structures. *Biopolymers*, 39(3):455–64. 1.4.3

Vakser, I. A., Matar, O. G., and Lam, C. F. (1999). A systematic study of low-resolution recognition in protein–protein complexes. *Proc Natl Acad Sci U S A*, 96(15):8477–82. 1.4.2.2

Valdar, W. S. and Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, 42(1):108–24. 1.1.3

Valencia, A. and Pazos, F. (2003). Prediction of protein-protein interactions from evolutionary information. *Methods Biochem Anal*, 44:411–26. 1.3.2.2

Vasilescu, J., Guo, X., and Kast, J. (2004). Identification of protein-protein interactions using in vivo cross-linking and mass spectrometry. *Proteomics*, 4(12):3845–54. 1.3.1.1

Verdonk, M. L., Cole, J. C., Watson, P., Gillet, V., and Willett, P. (2001). SuperStar: improved knowledge-based interaction fields for protein binding sites. *J Mol Biol*, 307(3):841–59. 1.4.2.3

Vert, J.-P. (2001). Introduction to support vector machines and applications to computational biology. 2.5.1.3

Vogel, B. E., Minor, R. R., Freund, M., and Prockop, D. J. (1987). A point mutation in a type I procollagen gene converts glycine 748 of the alpha 1 chain to cysteine and destabilizes the triple helix in a lethal variant of osteogenesis imperfecta. *J Biol Chem*, 262(30):14737–44. 1.1.2

Wade, R. C., Clark, K. J., and Goodford, P. J. (1993). Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J Med Chem*, 36(1):140–7. 2.3

Wade, R. C. and Goodford, P. J. (1989). The role of hydrogen-bonds in drug binding. *Prog Clin Biol Res*, 289:433–44. 2.3

Wade, R. C. and Goodford, P. J. (1993). Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. ligand probe groups with the ability to form more than two hydrogen bonds. *J Med Chem*, 36(1):148–56. 2.3

Wang, T. and Wade, R. C. (2003). Implicit solvent models for flexible protein-protein docking by molecular dynamics simulation. *Proteins*, 50(1):158–69. 1.4.2.3

Warwicker, J. and Watson, H. C. (1982). Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. *J Mol Biol*, 157(4):671–9. 2.3

Wiehe, K., Pierce, B., Mintseris, J., Tong, W. W., Anderson, R., Chen, R., and Weng, Z. (2005). ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. *Proteins*, 60(2):207–13. 1.4.2.4

Xu, D., Tsai, C. J., and Nussinov, R. (1998). Mechanism and evolution of protein dimerization. *Protein Sci*, 7(3):533–44. 1.1.2, 1.1.3

Yan, C., Dobbs, D., and Honavar, V. (2004). A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, 20 Suppl 1:I371–I378. 1.3.2.2

Yuan, Z., Zhao, J., and Wang, Z.-X. (2003). Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng*, 16(2):109–14. 2.4.8

Zacharias, M. (2003). Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*, 12(6):1271–82. 1.4.3

Zacharias, M. (2005). ATTRACT: protein-protein docking in CAPRI using a reduced protein model. *Proteins*, 60(2):252–6. 1.4.3

Zhang, C., Liu, S., Zhou, H., and Zhou, Y. (2004). An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci*, 13(2):400–11. 1.4.2.3

Zhang, C., Vasmatzis, G., Cornette, J. L., and DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol*, 267(3):707–26. (document), 1.4.2.3, 2.4.6, 2.5, 3.3.3

Zimmermann, O. (2002). *Untersuchungen zur Vorhersage der nativen Orientierung von Protein-Komplexen mit Fourier-Korrelationsmethoden.* PhD thesis, Universität zu Köln. 1.4.2.2, 2.2.1, 2.4.1

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat, dass sie – abgesehen von der unten angegebenen Teilpublikation – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Dietmar Schomburg betreut worden.

Keine Teilpublikation(en).

Oliver Martin

1. Referent: Prof. Dr. D. Schomburg
2. Referent: Prof. Dr. R. Schrader
Disputation: 12. Juni 2006

Curriculum Vitae

Oliver Sven Martin

Persönliche Daten

geboren am 8.März 1974 in Kaiserslautern
ledig
Staatsangehörigkeit: deutsch

Universitäre Ausbildung

seit 04/2002	Doktorarbeit am Lehrstuhl von Prof. D. Schomburg, Institut für Biochemie, Universität zu Köln. Arbeitsgebiet: Bioinformatik Thema: *"Efficient comprehensive scoring of docked protein complexes - a machine learning approach"*
03/2001-11/2001	Forschungsstipendiat in der Arbeitsgruppe von Prof. Andrew Torda, Research School of Chemistry, The Australian National University, Canberra, Australien. Arbeit zum Thema "Protein Threading"
06/2000-02/2001	Diplomarbeit im Fach Chemie im Arbeitskreis von Prof. D. Schomburg am Institut für Biochemie der Universität zu Köln *"Entwicklung und Anwendung funktioneller Oberflächenrepräsentationen von Proteinen"* Abschlussnote 1.2
02/1997-06/2000	Studium der Chemie an der Universität zu Köln, Spezialisierung in den Bereichen Biochemie/Bioinformatik
09/1993-01/1997	Studium der Chemie an der Universität Kaiserslautern

Schulausbildung

06/1993	Allgemeine Hochschulreife (Durchschnittsnote: 1.4)
1984-1993	Staatl. Gymnasium an der Burgstrasse Kaiserslautern
1980-1984	Theodor-Heuss-Grundschule Kaiserslautern