Application of Models for Biomolecular Structure and Interactions to Understand and Influence Biological Function

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
(Dr. rer. nat.)
Fakultät für Chemie und Biowissenschaften
Karlsruher Institut für Technologie (KIT) - Universitätsbereich
genehmigte

DISSERTATION
von
Irene Meliciani
aus
Rom (Italien)

Dekan: Prof. Dr. Martin Bastmeyer
Referent: Prof. Dr. Stefan Bräse
Korreferent: Apl. Prof. Dr. Wolfgang Wenzel
Tag der mündlichen Prüfung: 19.10.2011

a mamma e papà

Application of Models for Biomolecular Structure and Interactions to Understand and Influence Biological Function

Irene Meliciani

Karlsruher Institut für Technologie

Institut für Nanotechnologie

Intitut für Organische Chemie

Zusammenfassung

Die Forschung in den Lebenswissenschaften hat in den letzten Jahrzehnten zu einem enormen Wissensgewinns bezüglich der Abläufe zellulärer Prozesse und ihrer Steuerung geführt. Die Steuerung biologischer Prozesse wird vielfach über Kontaktwechselwirkungen der daran beteiligten Biomoleküle bewältigt. Mit dieser Arbeit möchte ich zu einem besseren Verständnis biomolekularer Interaktionen beitragen.

Ein Teil dieser Arbeit beschäftigt sich mit der Untersuchung von Proteininteraktionen, wobei als Methoden für die Modellierung von Protein-Protein-Wechselwirkungen (POEM) und für die Protein-Liganden-Wechselwirkung (FLEXSCREEN) eingesetzt wurden. Es wurden die Interaktionen der Chemokinrezeptoren CCR3 und CXCR1 mit für den Entzündungsverlauf wichtigen Steuerungsmolekülen (Chemokinen) untersucht. Dazu wurde ein *in silico* Alanin-Scanning-Protokoll entwickelt, welches experimentelle Verfahren zur Untersuchung wichtiger Interaktionen in der Protein-Protein-Kontaktfläche ergänzt. Die Untersuchungen zielen auf die Entwicklung neuer Inhibitoren für Chemokinrezeptoren ab, um inflammatorische Krankheiten besser behandeln zu können.

In einer weiteren Untersuchung wurde versucht, neue Wirkstoffmoleküle für die Cannabinoidrezeptoren CB1 und CB2 zu entwickeln, die in den Huntington und Alzheimer Krankheiten eine wichtige Rolle spielen. Dazu wurden Strukturmodelle für die Proteinrezeptoren erstellt und das Bindungsverhalten von Cumarinderivaten untersucht. In enger Zusammenarbeit mit dem Experiment wurde eine Reihe neuer Liganden synthetisiert und experimentell validiert, wobei sich eine erhöhte Affinität verglichen mit dem Referenzmolekül und eine gute, quantitative Übereinstimmung zwischen Simulation und Experiment zeigte. In einer letzten Anwendung im Bereich der rechnergestützten Medikamentenentwicklung wurde eine grosse Datenbank einem virtuellen Screening unterzogen, um neune Substanzen zu finden die regulierend in die Blutgerinnung einzugreifen.

Ein weiteres wichtiges Anwendungsfeld ist die Identifizierung genetischer Ursachen für Entwicklungsstörungen, die von patientenspezifischen Mutationen herrühren. Darüber hinaus habe ich Methoden der Protein-Strukturvorhersage angewandt, um die Relevanz von Mutationen, die an Patienten, die an Kallmann-Syndrom (KS), sowie normosmic hypogonadotropen hypogonadismus (nHH) leiden für den Krankheitsverlauf aufzuklären.

Diese Ergebnisse zeigen, dass rechnergestützte Verfahren zur Modellierung der Struktur von Biomolekülen und deren Wechselwirkungen heute dazu beitragen können, die komplexen zellulären Steuerungsmechanismen zu verstehen und gegebenenfalls auch zu beeinflussen. Damit erwächst aus den rechnergestützten Verfahren ein neues wichtiges Instrument um biologische Prozesse zu verstehen und die pharmazeutische Forschung voranzutreiben.

Table of contents

1. Overview

The completion of the human and the other genome projects has further accelerated the rapid growth of the available biological data. Efficient methods now exist to characterize the components of biological systems at the genetic, molecular, cellular and whole organism level, for example, the enzymes and metabolites in a metabolic pathway (Jeong, Tombor et al. 2000; Snoep, Bruggeman et al. 2006). All of these developments bring us closer to the tantalizing prospect to understand, control and ultimately design biological function. The recent progress of the genome sequence projects and of other molecular biology projects has increased the opportunities to seriously look into the possibility of system-level understanding.

One new promising branch is systems biology which has the potential to generate many new insights for biomedical research, industry and agriculture. Its primary goal is to understand life processes. Systems biology explores the dynamic life of processes at all levels, from the organization of cell organelles up to the complete cell, the genome via the proteome and to the entire organism. Methods from different fields are combined to explore these processes. Quantitative methods are now used in molecular biology incorporating knowledge from the fields of mathematics, informatics, and chemistry and physics. Systems biology focuses on the interplay of the molecular components that play an indispensable role forming symbiotic state of the system as a whole. Many areas are actively investigated, such as robustness of biological systems, network structures and dynamics, and applications to drug discovery (Kitano 2002; Kitano 2002; Kitano 2003).

One of the most important aspects to understand and control biological processes is the molecular level, where biological control is mediated by the interactions of specific biomolecules. Examples of such interactions are gene regulation (heat shock response, Fig. 1.1), protein-protein interaction (chemokine receptors in inflammation, Fig. 1.2) and small drug molecules (e.g. Cannabinoid receptor antagonists, Fig. 1.3). The heat-shock (HS) response is a good illustration of gene regulation where the heat-shock (HS) network ensures the survival of various organisms at different temperatures. Regulatory systems initiate a response that increases the resistance of cells to damage and aids in its repair, because they have evolved to detect the damage associated with stressors. HS response (Gross 1996) is one of the most important example of these systems. The HS response comprises elaborate mechanisms for detecting the presence of heat or other stressor-related protein damage (Ang, Liberek et al. 1991) and for initiating a response through the synthesis of new heat-shock proteins (HSPs) whose main function is to refold denatured cellular proteins.

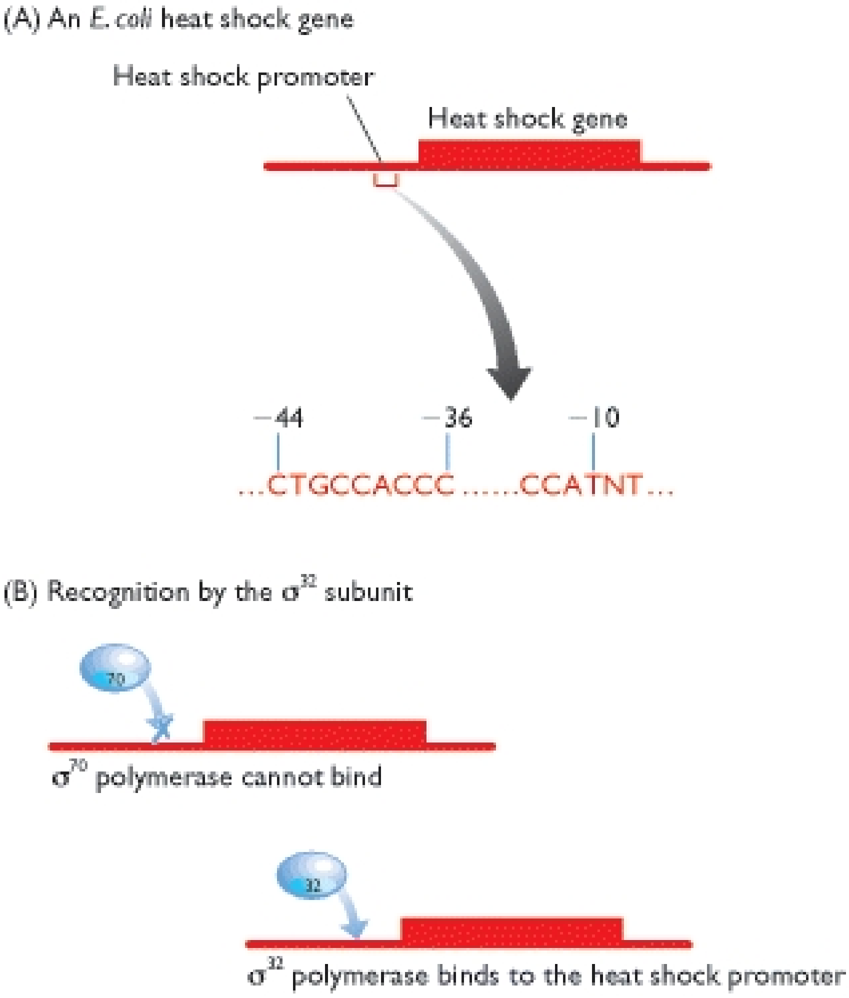

(A) An E. coli heat shock gene
Heat shock promoter
Heat shock gene
−44
−36
−10
...CTGCCACCC......CCATNT...
(B) Recognition by the σ32 subunit
70
σ70 polymerase cannot bind
32
σ32 polymerase binds to the heat shock promoter

Figure 1.1: example of the recognition of an *Escherichia coli* heat shock gene by the σ^{32} subunit (Brown 2002). (A) The sequence of the heat-shock promoter is different from that of the normal *E. coli* promoter. (B) The heat-shock promoter is not recognized by the normal *E. coli* RNA polymerase containing the σ^{70} subunit, but is recognized by the σ^{32} RNA polymerase that is active during heat shock.

In this thesis I have focused the theoretical investigation of biomolecular interactions with applications in drug discovery and the modulation of protein-protein interactions. Because such studies rely on the existence of a model for the protein receptor, I have also applied methods of protein structure prediction. Identification, quantification, and control over biological complexes are a key to research in drug-discovery, cell signaling, and elucidation of biosynthetic pathways and understanding of enzyme catalysis. Many of the projects discussed below combine experimental and computational approaches to better understand the nature of biomolecular interactions.

Protein-protein interactions are important in many biological networks, e.g. chemokine receptors play an important role in inflammation. One example is the leukocyte activation in acute inflammation. To reach sites of inflammation or injury, circulating leukocytes must exit the bloodstream by traversing the endothelium. The first step in the process of leukocyte recruitment at sites of inflammation is to bind and roll leukocytes on the endothelial cell surface by the generation of transient selecting-mediated interactions (Lawrence and Springer 1991). The slow velocity of rolling leukocytes allow a easier encounter with chemokines that are presented on the apical surface of the endothelium by glycosaminoglycans (Tanaka, Adams et al. 1993). Chemokines, which are discussed in detail in chapters 4.3-4.5 bind to their respective chemokine receptors expressed on the leukocyte cell surface, leading to the alteration of β2 integrin ambition on the leukocyte cell surface (Springer 1994). Then β2 integrins bind to their Ig counterligands, such as ICAM-1, ICAM-2, and ICAM-3, which have been up-regulated on the endothelial cell surface by proinflammatory cytokines. These interactions contribute a solid attachment of leukocytes to the endothelium and facilitate leukocyte migration (Rot 1993). The start of series of cellular events initiate by the binding of chemokines to their respective leukocyte receptors, such as changes in cell shape leading to enhanced locomotion, secretion of lysosomal enzymes, and production of superoxide anions.

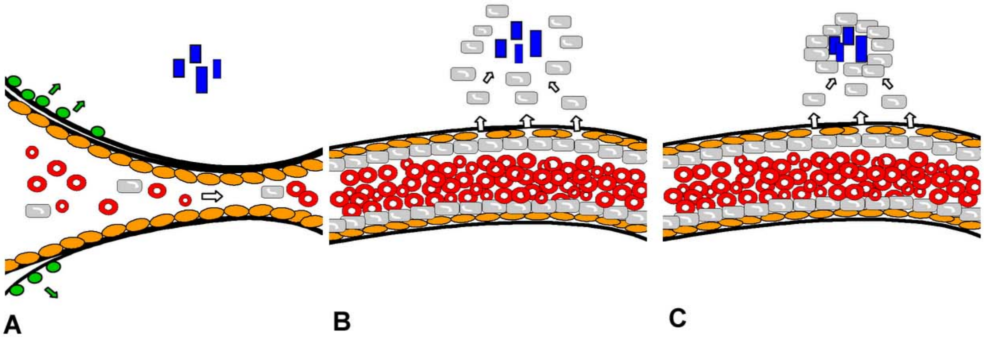


Figure 1.2: example of inflammation process. (A) Process of inflammation induce vasodilation , where mediator molecules alter the blood vessels to permit the migration of leukocytes, mainly neutrophils, outside of the blood vessels into the tissue. (B) neutrophils migrate along a chemotactic gradient created by the local cells to reach the site of injury. (C) binding of chemokines to their respective leukocyte receptors (taken from http://pocketstudy.blogspot.com/2007/04/infiammazione.html).

In this thesis I have cooperated with Dr. Katja Schmitz, the team leader of an experimental group to design molecules that influence the interaction between chemokines and their receptors CXCR1 and CCR3 (Meliciani, Klenin et al. 2009). This work was extended to predict the effect of all possible mutations on the previously identified hotspots of the extracellular domain of receptor CCR3 and CXCR1 that are crucial for chemokine binding. Elucidating the full binding motifs and locating their position on a three-dimensional model of the receptor reduces the experimental effort to investigate chemokine receptor recognition and activation.

This work was further extended to study small molecule binding to chemokine Interleukin-8. I performed *in silico* mutations, as for CCR3, on the N-terminal of the CXCR1 receptor, and investigated the binding energy of a set of 19 small molecules, which were active in a chip assay against chemokine interleukin-8 using receptor-ligand binding simulations. From the assay data it remained uncertain, whether the molecules would bind specifically or non-specifically to IL8. Only molecules binding at or near the receptor binding site IL8 are likely to interfere with chemokine receptor binding. The aims of this study was therefore to identify possible binding sites of the molecules with IL8 and to rank the small-molecule interactions for the different candidate binding sites. I found that four molecules had a significantly better binding energy than the average of all molecules and could be tested experimentally with one of the predicted binding site that is near the position where interleukin-8 interacts with his receptor.

One important application of investigations targeting biomolecular interactions is drug design, which we later address in the development of CB1 receptor antagonists (chapter 5.4) and other applications (chapters 5.3 and 5.5). The first antagonist for cannabinoid receptor was rimonabant (Rinaldi-Carmona, Barth et al. 1994; Barth and Rinaldi-Carmona 1999). Rimonabant is a selective cannabinoid CB1 receptor antagonist that has undergone extensive testing in the treatment of obesity, as well as for treating nicotine dependence in humans. The endocannabinoid system interacts with several neuropeptides that modulate hunger and satiety signals, resulting is the stimulation of appetite (Di Marzo and Matias 2005). The cannabinoid receptor, CB1, was found extensively in the brain (Matsuda, Lolait et al. 1990). CB1 receptors appear to regulate the activity of mesolimbic dopamine neurons, (Spanagel and Weiss 1999) and to interact with neuropeptides such as the melanocortins and gut peptides such as ghrelin in regulating food intake (Tucci, Rogers et al. 2004; Verty, McFarlane et al. 2004). This discovery led to the development of numerous CB1 antagonists, of which rimonabant is having the most success in human treatment.

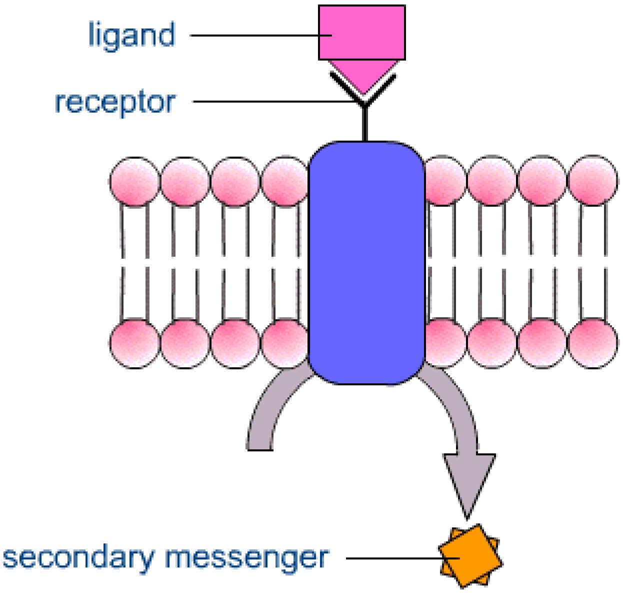


Figure 1.3: Contact between ligand and receptor activate a secondary messenger (taken from http://www.galenotech.org/farmcodin2bis.htm).

In this direction I have participated in an investigation and optimization of coumarin-derivatives to cannabinoid receptor CB1 and CB2. The cannabinoid receptors belong to membrane-bound G protein-coupled receptor superfamily and are predominantly coupled via Gi/o proteins. Many factors, including receptor conformational change, differences in long-range interactions and local factors contribute to receptor selectivity.

The CB1 receptor subtype is known to play an essential role in analgesia, memory-impairment, spasmolysis and regulation of appetite. Selective agonists for CB2 receptors show antiinflammatory and analgesic properties in animal models.

Because no crystal structures are available I developed homology models and performed small-molecule docking studies for cannabinoid receptors in cooperation with the group of Prof Bräse at KIT and the group of Prof. Müller from PharmaCenter Bonn. I succeeded in predicting modifications of previously known small-molecule ligands for the CB1/CB2 receptor pair, which not only exhibited higher affinity in subsequent experiments, but which also demonstrated the receptor selectivity between CB1/CB2, as observed in the simulations (Congress 2011). The primary goal of the simulations was to understand changes in the binding mode and the binding energy between the new compounds and structurally related compounds with a known affinity. This work resulted in the discovery of new ligands with aliphatic sidegroups and is also a successful example using an homology modeling for the investigation of protein-ligand interactions.

In a related project I have participated in an extended cooperation to develop new molecules that modulate thrombin activation by antithrombin. The goal of the project is the development of a structure-based model for AT-II regulation to develop new anticoagulation drugs. We performed docking simulations for the complex Antithrombin/Heparin using FlexScreen, which reproduced the known binding mode very well. In the second part of the project we performed a large-scale *in silico* screen and compared the binding scores with those of heparin. The compounds with the best energy values have been tested *in vitro* by our collaborating group in Murcia (Spain), where we found out that one of the compounds increases heparin binding with antithrombin and thrombin activation (patent pending).

In February 2008 my group and I were approached by the groups of L. Layman and J. Gusella from Massachusetts General Hospital/Harvard Medical School to assist their investigation to identify the genetic causes of human disease, involving a series of genes causing Kallmann syndromes (KS) and Normosmic hypogonadotropic hypogonadism (nHH). To support this work I have applied techniques to generate the models for the three-dimensional structure protein, for which no experimental structure could be determined to date, from the amino acid sequence. I have mainly used techniques in comparative structure prediction (also called homology modeling), which is based on the assumption that two homologous proteins will share very similar structures. Because a protein's fold is more evolutionarily conserved than its amino acid sequence, a target sequence can be modeled with reasonable accuracy even on only distantly related templates, provided that the relationship between target and template can be discerned through sequence alignment. I performed studies on protein structure prediction for *CHD7*, *WDR11* and *NELF* to analyze possible functional effects of the mutations in the patients exhibiting the syndrome (Kim, Kurth et al. 2008; Kim, Ahn et al. 2010; Xu, Kim et al. 2011).

This thesis is structured as follows:

- chapter 2 gives a general introduction to the state of the art related to this thesis;
- chapter 3 is focused on the application of protein structure prediction to identify genetic causes for human disease;
- chapter 4 reports the work on protein-protein interactions, specifically chemokine-receptor interactions;
- chapter 5 reports the work on protein-ligand interactions, specifically coumarin derivatives-CB1/CB2 receptor interactions, and the discovery of new molecules affecting blood coagulation;
- chapter 6 gives a short summary and an outlook to future developments.

2. Introduction

2.1. Introduction

Life science research in the last decades has lead to an enormous increase in the amount of available data, encompassing all levels of biological organization, starting at the genome and ending at the whole organism (Rogers 2000). Much of this data is hard won in experimental efforts that shed light on specific biological mechanisms, generating islands of knowledge in a sea of the unknown. Bioinformatics is an integrating effort that attempts to link biological data using methods of computer science generating computational methods that are used as a support for biology studies. In addition to modeling and simulation, bioinformatics attempts to describe biological events from a numerical and statistical point of view. The term bioinformatics was used for the first time by Paulien Hogeweg in 1979 (Hogeweg 1978), at first for using bioinformatics in genomics for studies in DNA sequencing and much effort has since been invested to develop methods for many fields, including systems biology. Bioinformatics is now also covering other areas of investigation, including sequence alignment, gene prediction, protein structure prediction, biomolecular interactions, drug design, drug discovery, protein structure alignment and many others.

The goal of bioinformatics is the development of approaches to collect, link and structure biological data, e.g. in specific databases, such as the Protein Data Bank that 'contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies' (Rose, Beran et al. 2011) or creating algorithms with the objective to contribute to the comprehension of molecular mechanism in biology. The recent expansion of the internet allowed globalization of a much biological data leading to development of many databases and different bioinformatics programs, such as Phyre (Kelley and Sternberg 2009) for protein structure prediction.

Bioinformatics is destined to continue to grow and support more and more experimental lines of research and discovery, possibly aiding in the discovery of the causes of diseases or the change of organisms to the environment, for example by exhibiting specific traits using single point mutations (Ng and Henikoff 2003; Flanagan 2010). Such development could, for example, help prevent and treat diseases, improve food production, and preserve of the environment.

In this chapter, I will first give a brief overview of protein structure (section 2.2), followed by a synopsis of the most important biophysical/biochemical interaction mechanisms that govern protein conformational change and function (section 2.3). In section 2.4 I will give an overview of the general state-of-the-art of methods for protein structure prediction, which is the basis of the theoretical work reported in chapter 3. This is followed by a summary of methods treating protein-small-molecule interactions in section 2.5. Within this wide field special emphasis is given to docking methods, as these have been applied for the studies in chapter 4 and 5. I conclude this chapter with a summary of extensions of these methods studying protein-protein interactions, which is the background for the work reported in chapter 4. Given the wide range of topics covered the review of methodology cannot be comprehensive but is focused to provide the background material for the studies reported in subsequent chapters. Details of the methodology applied in the specific protocols are given in the methods section of the corresponding chapters.

2.2. Protein Structure

Proteins among the most important biological macromolecules in all cells, comprising of one or more polypeptide chains, composed of a sequence of the twenty amino acids occurring in nature. Understanding protein function, control and interactions is therefore one of the most important goals in biology. Proteins often fold into unique three dimensional structures and knowledge of these structures is often essential to understand protein function. The building blocks of proteins are amino acids, which comprise an amino and a carboxylic acid group and are interlinked by peptide bonds in the polypeptide chain. The individual amino acids differ by their sidechain (designated R in the Fig. 2.1). The sidechains of the naturally occurring proteins are listed in figure 2.1. Protein synthesis takes place in the ribosome, to form very complex structures. The structure of proteins has traditionally been classified into four categories:

Primary Structure: Primary structure labels the sequence of amino acids linked by a peptide bond that starts always from the amino (N) terminus to carbonyl (C) terminus in one polypeptide chain. It labels the logical, rather than the physical organization of the polypeptide chain. The sequence of the chain is written in either one-letter code or three-letter code, representing the 20 naturally occurring amino acids (Fig. 2.1).

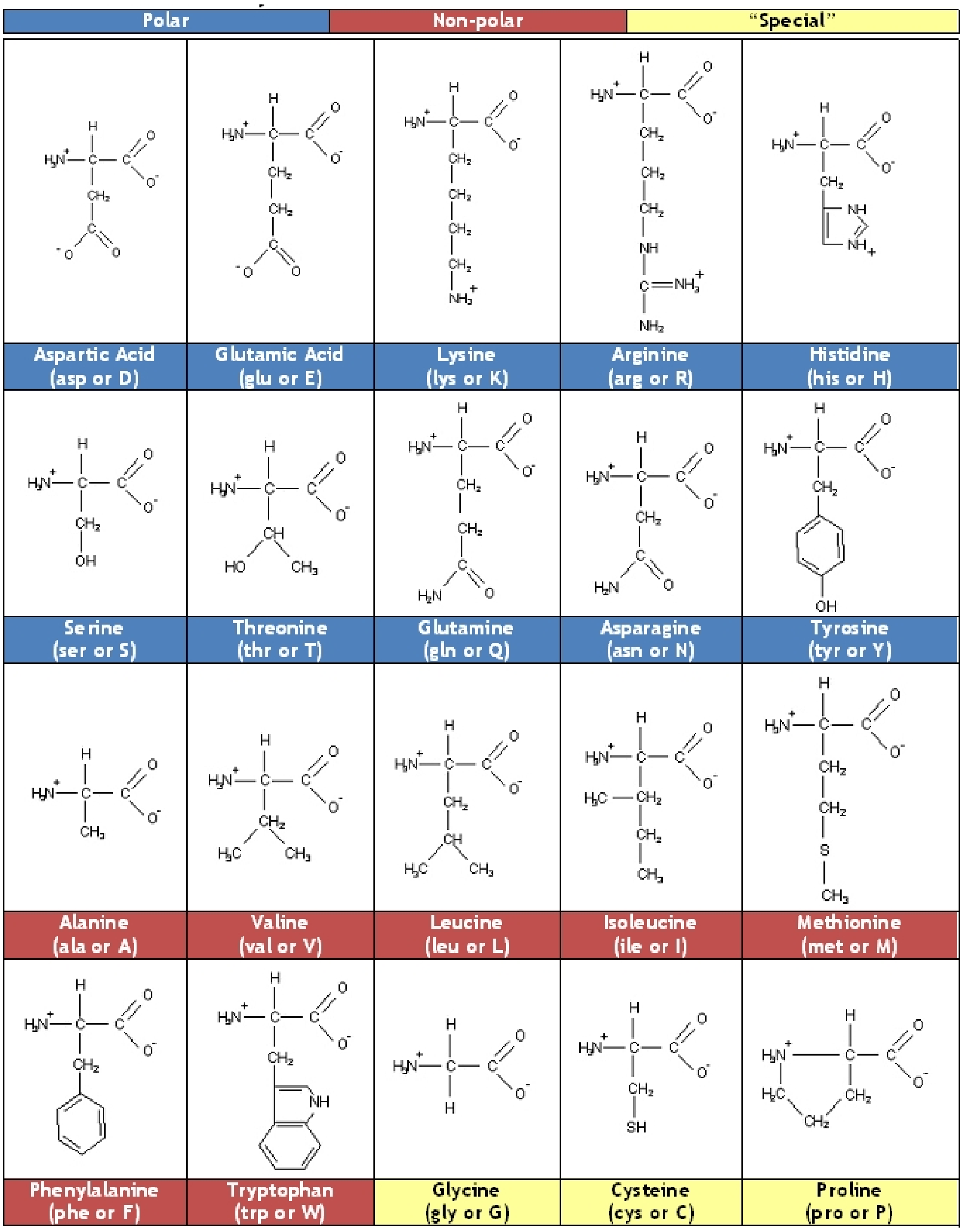


Figure 2.1: List of the 20 amino acids (taken from http://medchrome.com, Amino acids: Memorizing names and classifying made easy).

Secondary Structure: Secondary structure labels recurring local conformations that the chains can assume based on the amino acids sequence and formation of hydrogen bonds between the main peptide chain. These bonds result in a specific geometry due to the constraints of the dihedral angles ψ and φ as illustrated by the Ramachandran plot (Fig. 2.2). In fact not all the values of angles ψ and φ are possible, because many (ψ / φ) combinations produce too close contacts between the atoms. The Ramachandran plot (Ramachandran, Ramakrishnan et al. 1963) shows three main areas that are energetically favorable, which correspond to the most widely observed secondary structure elements identified by L. Pauling in 1951 (Pauling and Corey 1951), e.g. α-helix and β-sheet. All secondary structure elements feature hydrogen bonds between the NH and CO groups of the main chain (Fig. 2.3).

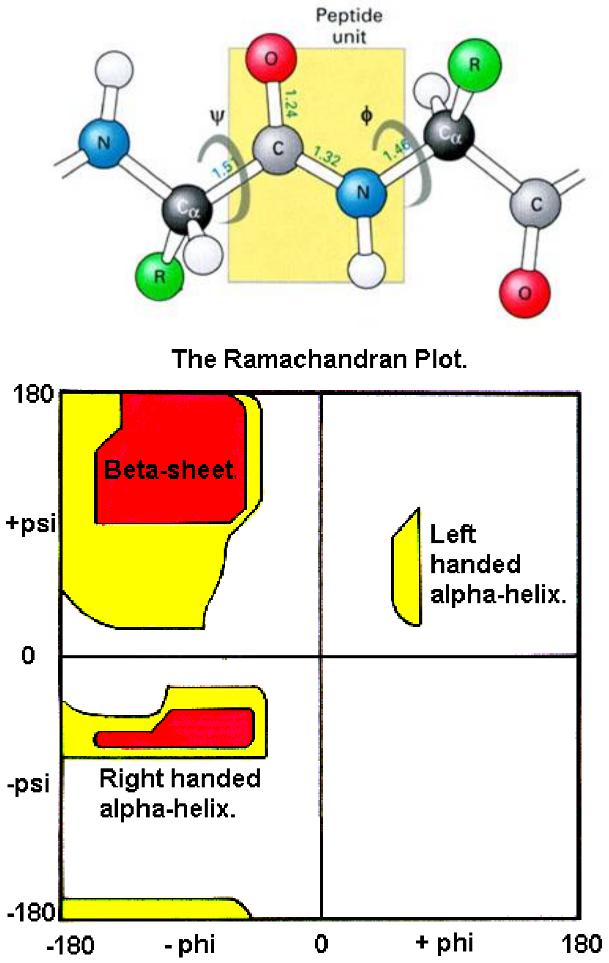


Figure 2.2: (A) Schematic representation of an amino acid in a polypeptide chain, indicating the dihedral angles which are the only low-energy degrees of freedom of the backbone of a protein and (B) ramachandran plot indicating permitted values of the dihedral angles and the corresponding secondary structure due to the constrains of the dihedral angles (Ramachandran, Ramakrishnan et al. 1963).

The α-helix is the most common conformation adopted by the polypeptide chain, which forms a spring-like structure where the tightly coiled backbone forms the inner part of the helix and the side chains project outwards in a helical array. In the α-helix the CO group of each amino acid forms a hydrogen bond with the NH group of the amino acid which is situated four residues ahead in sequence; except for the terminal residues in the helix, all the carbonyl and amine groups are hydrogen bonded.

In an α-helix axis amino acids are spaced 1.5 Å along the helix direction, which results in 3.6 amino acids per turn of the helix. Both right-handed and left-handed helices can occur, but most of the observed α-helices in naturally occurring proteins are right-handed. There are also other types of helices such as 3_{10}-helix, pi-helix, polyproline II helix and collagen helix. All these helices are classified by their hydrogen bonding pattern. The polyproline II and collagen helix are left-handed, while other helices are right-handed.

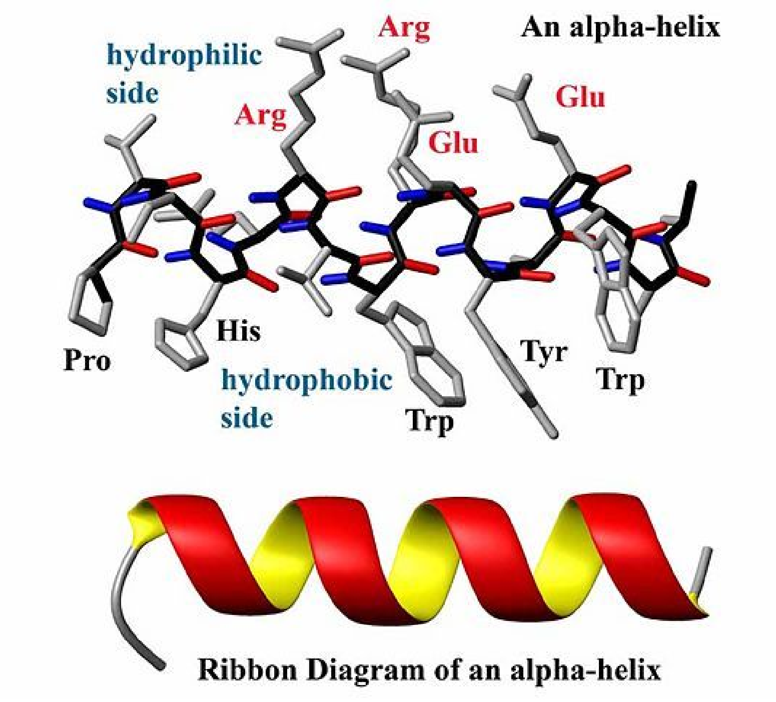


Figure 2.3: Structure and ribbon diagram of a protein alpha helix. The backbone atoms form a coil (black bonds) while the carbonyl groups (red) form hydrogen bonds with the amide groups (blue) (taken from http://en.citizendium.org/wiki/Protein_structure).

The structure of β-sheets differs remarkably from the spring-like structure of the α-helix. β-sheets, see (Fig. 2.4), consist of two extended chains, called β strands. A β sheet is formed by linking two or more β strands by hydrogen bonds (Kabsch and Sander 1983; Richards and Kundrot 1988; Frishman and Argos 1995).

The strands are spaced with 3.5 Å distance between adjacent amino acids (this is in contrast to the 1.5 Å of the α-helix), where adjacent amino acids have the side chains pointing in opposite directions. Parallel, anti-parallel or mixed β-sheets occur in nature: In the anti-parallel β-sheet adjacent chains run in opposite direction, the NH group and the CO group of each amino acid are bonded to the CO and NH group of a partner on the adjacent chain, respectively. In the parallel β-sheet adjacent chains run in the same direction, the NH group is hydrogen bonded to the CO group of one amino acid on the adjacent strand, whereas the CO group is hydrogen bonded to the NH group on the amino acid two residues further along the chain.

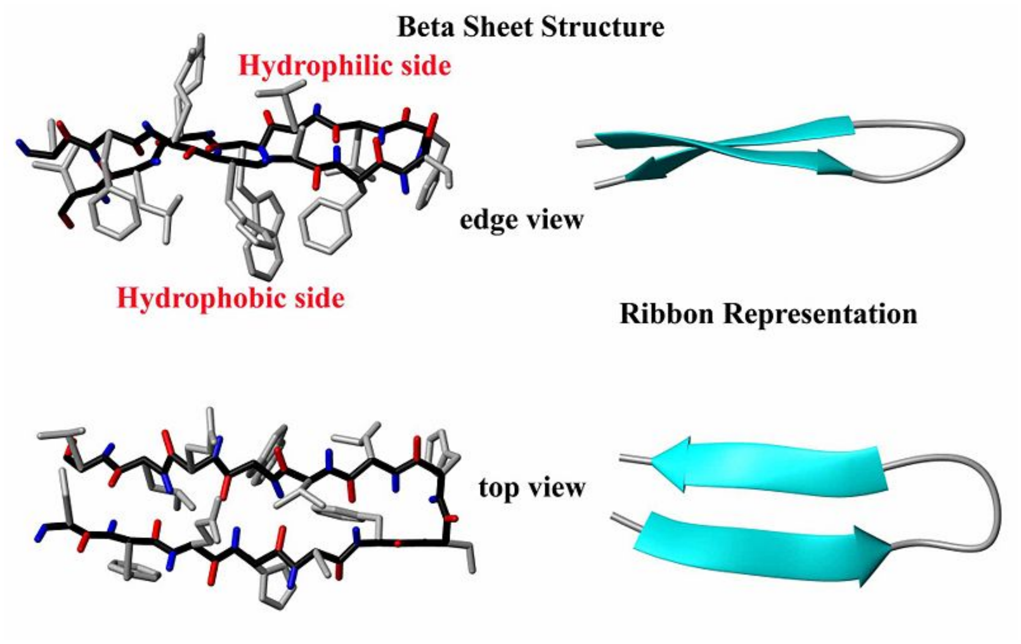


Figure 2.4: Structure and ribbon diagram of a protein beta sheet. Note the oscillating positions of the carboxyl (red) and amide (blue) groups which form hydrogen bonds between the two beta strands (taken from http://en.citizendium.org/wiki/Protein_structure).

Tertiary Structure: The tertiary structure of a protein refers to the three-dimensional arrangement of the secondary structure elements in space (Fig. 2.5). It is formed by the folding of secondary structural elements such as α-helix, β-sheets, loops and turns. The overall arrangement of the atoms in space, also called the native structure, is responsible for the biological function of the protein. Many proteins fold spontaneously into this native structure, driven by a number of biophysical/biochemical interactions. In addition to hydrogen bonding the hydrophobic effect is important determining the structure of globular proteins, many of which are characterized by a hydrophobic core and surface charged/polar residues.

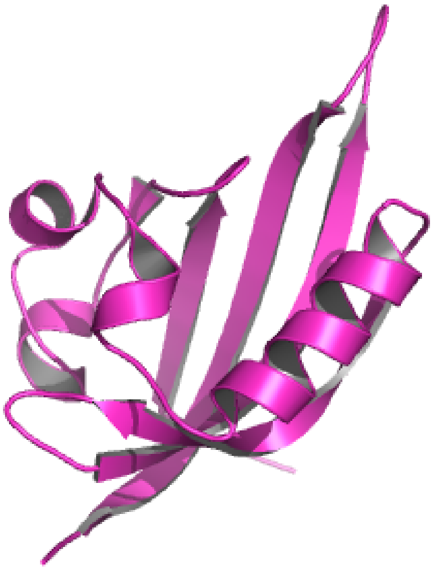


Figure 2.5: example of tertiary structure, Endothelial PAS domain protein 1 (PDB code 1P97).

Quaternary Structure: The term quaternary protein structure refers to the association of two or more polypeptide chains to form a complex (Fig. 2.6). Multimeric complexes are important for various biological activities, as for example enzyme activity when residues from more than one polypeptide subunit are forming the active site, or when adjacent active sites may be involved sequentially in catalysis of a complex reaction.

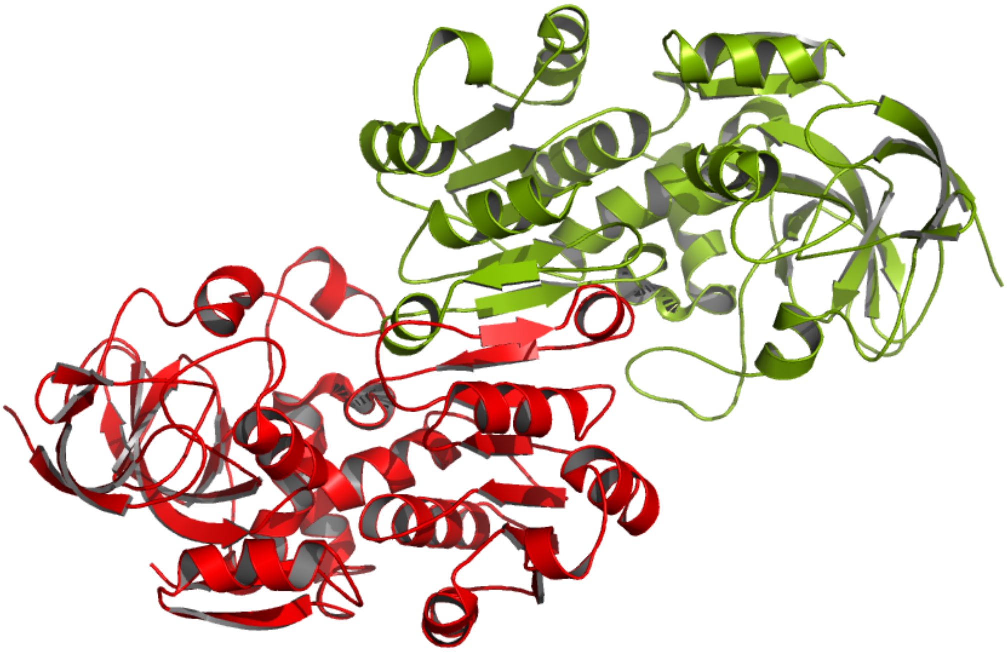


Figure 2.6: example of a dimer the simplest possible quaternary structure. Glutathione-dependent formaldehyde dehydrogenase (PDB code 1M6H).

2.3. Protein Structure Prediction

The 3-D structure of proteins aids in clarifying their properties, behavior and almost all biological phenomena mediated by proteins, including protein-ligand, protein-protein interactions, drug function and protein design. Experimental efforts have led to deposition of more than 53,000 experimentally solved structures in the Protein Data Bank (PDB), but there are far more protein sequences reported, with >400,000 in Swiss-Prot and >7,500,000 in TrEMBL (Jain, Bairoch et al. 2009) (Fig. 2.7). Protein structure prediction remains a significant challenge because the rapidly increasing number of reported protein sequences is not matched by experimental techniques to solve all the associated structures.

Structure prediction methods (Ginalski 2006; Xiang 2006) represent one promising route to bridge the gap between sequence and structure. This objective is realistic considering that sequences with 50% or more identity can often be modeled within experimental accuracy (Kryshtafovych, Fidelis et al. 2007). Structure prediction methods are classified into homology modeling (HM), also called template-based modeling (TBM) or comparative modeling (CM), and free modeling (FM) (Zhang 2008). Proteins with similar sequence fold into similar 3-D structures. In HM, the 3-D structure of the protein is built from a template structure, but this works only if a protein with sufficiently high homology and solved structure is available. In the absence of this information FM are used, which do not depend on a priori structural information. The success rate of FM methods is presently low (Peng and Xu 2010).

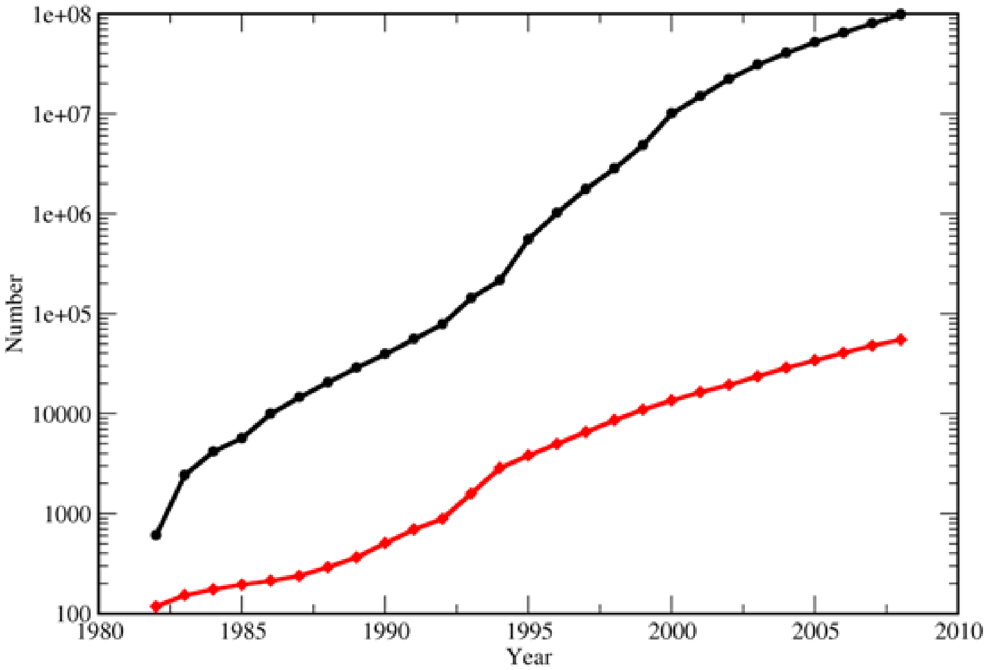


Figure 2.7: Growth of Biological Databases. <u>Black line:</u> protein sequences deposited in TrEMBL (protein sequence database) <u>Red line:</u> proteins whose 3-D structure has been solved 75,594 (PDB) protein structures (30/08/2011).

2.3.1. Comparative Modeling

Many different methods for structure prediction have been proposed (Zaki and Bystroff 2008), but the most widely used methodology is comparative structure prediction (also called homology modeling), where the model of the "target" protein is constructed from the amino acid primary structure and from an experimental three-dimensional structure of a related homologous protein called "template" (Fig. 2.8). Homology modeling capitalizes on the fact that in nature there is a limited number of folding conformations (Chothia 1992), as a result the three-dimensional protein structure is evolutionarily more conserved.

In homology modeling we have to find one or more known structures that are similar to the structure of the query sequence and then produce an alignment between the residues of the query sequence and the residues in the template sequence (Honig, Tang et al. 2003; Elofsson, Ohlson et al. 2004; Marti-Renom, Madhusudhan et al. 2004). The target model is created from the sequence alignment and the template structure. The higher the quality of the alignment between the sequence and the template is, the better the model that is produced. Gaps in the sequence alignment or gaps in the template structure can decrease the quality of the model. Loops region are normally the most involved in this kind of errors, where the target and template proteins may be completely different.

When there is no high-homology template for the whole sequence, we have to use loop modeling techniques to fill the gaps in the structure, which typically result in less accurate model structures. Low quality models can typically not be used for studies such as drug design and protein-protein interaction prediction; but sometimes they can still be used to obtain some interesting information of the biochemistry of the query sequence, for example by making hypotheses about the conservation of certain residues, information that could also be useful for experimentalist.

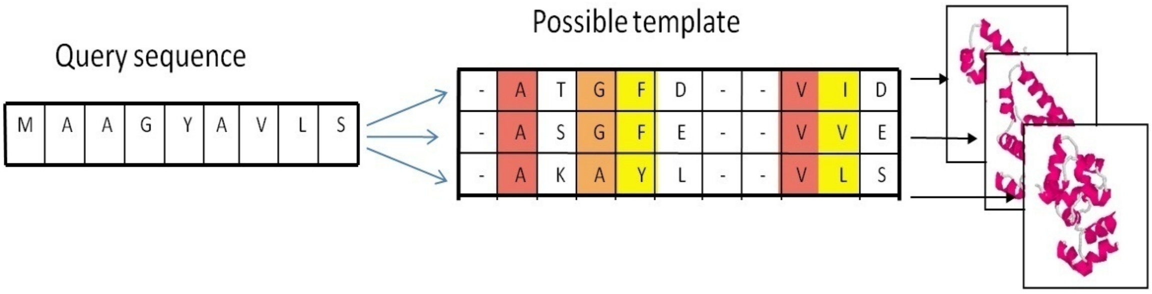


Figure 2.8: Homology modeling procedure, from the query sequence we have to find one or more known structures called template to generate the model.

2.3.2. Model Building

Given an alignment there are four principal methods for model construction, including the spatial restraint method (SSR) (Sali and Blundell 1993), the segment matching method (SMM) (Levitt 1992), the multiple template method (MTM) (Chothia, Lesk et al. 1986; Blundell, Sibanda et al. 1987) and artificial evolution (AE) (Petrey, Xiang et al. 2003). SSR assumes that several geometrical features such as distances and angles are conserved in homologous proteins, when comparing equivalent positions. Two main steps are involved, extraction of spatial restraints based on alignment and construction of the target 3-D model by fulfilling the spatial restraints (Sali 1995). One of the most used frequently homology modeling programs is MODELLER that uses SMM to divide the target into a series of short segments, each matched to its own template fitted from the PDB. The alignment of the sequence is done over segments rather than over the entire protein. There are different steps in this method including the construction of a segment database, a model construction via iterative randomization to get an average model and a minimization to get the final model. An extension to the SMM program SegMod/ ENCAD called Pfrag was proposed by Larsson et al. which can use multiple templates (Larsson, Wallner et al. 2008).

In MTM, several solved protein 3-D structures are used to build the target protein model. Based on sequences and structures the multiple templates are aligned with each other and the target is optimally aligned with the multiple templates. Loops are present between the conserved regions where they are usually exposed at the surface of the proteins. MTM has been implemented in several packages such as SWISS-MODEL (Schwede, Kopp et al. 2003) and MOE (Boyd 2005), 3D-JIGSAW (Bates, Kelley et al. 2001).

In AE, the alignment of the sequences of the template and target is achieved using evolutionary concepts like mutations, insertions and deletions. The model of the target protein is built by editing the template structure based on the alignment. The aligned residues are mutated and this causes changes in scoring function (energy) which is minimized in the procedure. Algorithms for modeling mutations are followed by procedures modeling deletions and then insertions. When there is no a significant energy penalty the operation is considered successful (Petrey, Xiang et al. 2003).

Some automated web servers such as I-TASSER (Skolnick, Zhou et al. 2009), ROBETTA (Baker, Raman et al. 2009), Pmodeller-6 (Wallner, Larsson et al. 2007) are implemented in consensus servers, such as Pcons (Wallner, Larsson et al. 2007), and HHPred3 (Soding 2005), which have been very successful in accurate prediction of the protein target structures. I-TASSER searches the whole PDB library to find appropriate protein fragments from which the global structure is assembled by combining aligned fragments. For portions for which no alignment matches are found, the 3-D structure is built using *de novo* simulations. The final refinement of the model is made with a search of the lowest energy conformation (Zhang 2007; Zhang 2008).

Model building in Pcons is carried out using Pfrag (Larsson, Wallner et al. 2008), a modified SegMod homology modeling program, and final refinement is performed using the ENCAD forcefield (Wallner, Larsson et al. 2007). Model prediction by ROBETTA makes use of extensive and computationally expensive conformational sampling and all-atom energy refinement (Chivian, Kim et al. 2003). Others web servers are M4T (Multiple Mapping Method with Multiple Templates) (Fernandez-Fuentes, Madrid-Aliste et al. 2007) and PROTEUS2 (Montgomerie, Cruz et al. 2008).

2.3.3. Loop Modeling

Loop modeling methods can be classified into two major approaches: (i) knowledge based and (ii) energy based methods. A few methods have also been reported which combine the two approaches. Knowledge based methods are limited by the availability of relevant loop structures from known protein structures (Peng and Yang 2007) because it is difficult to find a suitable loop segment that fits between the two stem regions of the loop from a database of structures. A database of structural motifs, ArchDB, was developed (Espadaler, Fernandez-Fuentes et al. 2004) and evaluated using two different sequence profiles. Other methods use Monte-Carlo simulation of the loop, ranking the fragment database for the loop prediction using the DFIRE potential (Lee, Seok et al. 2008). Energy based methods use a *de novo* energy function for conformational search of the loops to test their quality (Soto, Fasnacht et al. 2008). *De novo* loop modeling is a good method for loops not longer than seven residues (Jacobson, Pincus et al. 2004). Loop conformational search can be carried out using tools such as local move Monte-Carlo (LMMC) (Cui, Mezei et al. 2008), torsion angle conformational search (Felts, Gallicchio et al. 2008), LoopBuilder (Xiang, Soto et al. 2002), replica exchange (Olson, Feig et al. 2008) or a dihedral angle-based buildup procedure in hierarchical loop prediction (HLP) (Jacobson, Pincus et al. 2004). The conformers generated are scored using a forcefield or other physics-based energy calculations (Zhu, Pincus et al. 2006) usually including solvation effects. LOOPER allows a systematic and efficient sampling strategy, searching for loop conformers with optimal interactions of the loop backbone with the rest of the protein atoms. For the final ranking the CHARMm energy scoring function with a generalized Born solvation term is used (Spassov, Flook et al. 2008).

2.3.4. Side Chain Modeling

Regarding side chain prediction most of the methods use rotamer libraries which are constructed using statistical knowledge of protein 3-D structures such as the Grow-to-Fit molecular dynamics method (G2FMD) (Zhang and Duan 2006), statistical machine learning methods (Yan, Kloczkowski et al. 2007) and IRECS that selects more than one rotamer in order to have a representation of the conformational space flexibility of the side-chain. The ranking is calculated by a knowledge based statistical potential, ROTA (Hartmann, Antes et al. 2007).

To improve side-chain modeling modifications to the ROSETTA energy functions with softer van der Waals terms and extended rotamer libraries have been implemented (Dantas, Corrent et al. 2007). A novel search method and novel energy function were proposed to predict global minimum more reliably (Yanover, Schueler-Furman et al. 2008), based on tree reweighted belief propagation. Another method, OPUS-Rota, combines the recently OPUS-PSP potential with a Monte-Carlo conformational search (Lu, Dousis et al. 2008).

2.3.5. Quality Assessment

Quality assessment (QA) methods attempt to rate the quality of the model based on statistical evidence. There are statistical as well as physico-chemical methods, which are based on alignment to a single template or multiple templates or on meta server results. The QA method gives a local score as a function of residue or residue window (Wallner and Elofsson 2006; Gao, Bu et al. 2007; Wiederstein and Sippl 2007; McGuffin 2008; Mereghetti, Ganadu et al. 2008) or a global score (Eramian, Shen et al. 2006; Benkert, Tosatto et al. 2008; Qiu, Sheffler et al. 2008; Randall and Baldi 2008) which may be based on single or multiple assessment criteria. Some programs for quality assessment are ModFOLD (McGuffin 2008) server that combines ModSSEA (McGuffin 2007), MODCHECK (Pettitt, McGuffin et al. 2005) and ProQ (Wallner, Fang et al. 2003), AIDE (Mereghetti, Ganadu et al. 2008) scores with secondary structure information. The local quality of a structure can be quantified using ProQres, which relies on 3-D information, or ProQprof, using a model generated from sequence alignment (Wallner and Elofsson 2006). ProQres quantifies structural qualities such as secondary structure, solvent accessibility, and atom-atom and residue-residue contacts to have a measure of local quality. ProQprof uses profiles both for target and template (the sum of the two scores is arranged as ProQlocal). ModelEvaluator quantifies the absolute quality of a protein model using support vector regression (SVR) (Wang, Tegge et al. 2009), and was trained using only structural characteristics such as secondary structure, contact map, relative solvent accessibility and β-sheet structure.

2.3.6. State of the Art

Advancements in these efforts are monitored in the CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiment. CASP is a series of protein structure prediction community experiments conducted every two years since 1994. Many groups working on methods in structure prediction submit the results of their predictions for a set of targets whose experimental structures are available but not yet revealed. CASP has over the years become a good indicator of the progress of protein based structure prediction methodologies. From CASP 6 to 7 improvements have been made mainly for medium or high difficulty targets (Kryshtafovych, Fidelis et al. 2007), a trend that continued in CASP8. Accuracy and modeling of regions not available from the template (Battey, Kopp et al. 2007;

Kryshtafovych, Fidelis et al. 2007; Kryshtafovych, Fidelis et al. 2009) showed improvements throughout the CASP experiments in particular have been improvements in the performance of fully automated servers. Best predictions were those that used human expertise (Battey, Kopp et al. 2007; Kryshtafovych, Fidelis et al. 2007) where the best six predicted structures by humans or automated servers in CASP7 and CASP8, ~29% were from the servers. This is a clear improvement over CASP5 and CASP6, in which the number was ~15%. In addition for 90% of the CASP8 targets at least one of the top six predictions was from an automated server (Battey, Kopp et al. 2007; Kryshtafovych, Fidelis et al. 2007; Kryshtafovych, Fidelis et al. 2009). This is a success particularly for large scale modeling approaches, where confidence on human expertise can be very expensive.

Many methods are available via web-services, a non-comprehensive summary is given in Table 2.I.

Name	Method	Website
3D-JIGSAW	Fragment assembly	http://bmm.cancerresearchuk.org/~3djigsaw/
Biskit	wraps external programs into automated workflow	http://biskit.pasteur.fr/
CABS	Reduced modeling tool	http://www.biocomp.chem.uw.edu.pl/services.php
CPHModel	Fragment assembly	http://www.cbs.dtu.dk/services/CPHmodels/
EasyModeller	GUI to MODELLER	http://sites.google.com/site/bioinformatikz/
ESyPred3D	Template detection, alignment, 3D modeling	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
FoldX	Energy calculations and protein design	https://genesilico.pl/meta2/
GeneSilico	Consensus template search/fragment assembly	https://genesilico.pl/meta2/
Geno3D	Satisfaction of spatial restraints	http://geno3d-pbil.ibcp.fr/cgi-bin/geno3d_automat.pl?page=/GENO3D/geno3d_home.html
HHpred	Template detection, alignment, 3D modeling	http://toolkit.tuebingen.mpg.de/hhpred
LOMETS	Local Meta threading server	http://zhanglab.ccmb.med.umich.edu/LOMETS/
MODELLER	Satisfaction of spatial restraints	http://salilab.org/modeller/
Phyre and Phyre2	Remote template detection, alignment, 3D modeling, multi-templates, *ab initio*	http://www.sbg.bio.ic.ac.uk/~phyre/
Protinfo CM	Comparative modeling of protein structure using minimum perturbation and loop building	http://protinfo.compbio.washington.edu/abcm/
ROBETTA	Rosetta homology modeling and ab initio fragment	http://robetta.org/

	assembly with Ginzu domain prediction	
Selvita Protein Modeling Platform	Package of tools for protein modeling	http://www.selvita.com/
SWISS-MODEL	Local similarity/fragment assembly	http://swissmodel.expasy.org/
TIP-STRUCTFAST	Automated Comparative Modeling	https://tip.eidogen-sertanty.com/Login.po
WHAT IF	Position specific rotamers	http://swift.cmbi.kun.nl/whatif/

Table 2.I: list of protein structure prediction software,
(taken from http://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software).

2.4. Protein Ligand Interactions

Protein-ligand interactions have a central role in all processes in living systems. Understanding of protein interactions with small molecules is of great interest as it allows to understand and influence protein function, also for therapeutic applications. Protein conformational changes, including folding, as well as many biological functions are mediated by intra- and intermolecular interactions. Models for such interactions are required to predict preferred orientation of the molecules and the strength of association or binding affinity between two molecules.

Among the most important non-covalent inter-atomic interactions mediating the binding of small molecules with protein are, electrostatic and van der Waals interactions (Fig. 2.9). Other important factors that contribute to protein-ligand affinity, include entropy changes in the solvent and intramolecular contributions arising from the flexibility of the receptor in the binding site (Gohlke and Klebe 2002; Bissantz, Kuhn et al. 2010). Hydrogen bonding, salt bridge and metal interactions (Gohlke and Klebe 2002) are often treated in the framework of the electrostatic model, but in particular for hydrogen bonding, one of the most important interactions in biological macromolecules, this may not be fully justified. Hydrophobic interactions involve contacts between non-polar parts of the molecule and have been shown to play a crucial role in ligand binding (Bissantz, Kuhn et al. 2010).

Such models may be used to model the process of receptor ligand binding in order to determine the binding pose and the binding affinity. However, due to computational constraints, is presently not possible to model the binding/unbinding process in unbiased molecular dynamics simulations, which would be a direct *in silico* equivalent of the experiment. Instead two families of methods have emerged that address different aspects of the problem: Docking methods, reviewed in section 2.4.1 aim the rapid prediction of binding modes and relative affinities for a set of potentially interesting compounds. Because many of these methods aim at high-throughput screening they use often simplified models, e.g. excluding receptor flexibility. These methods are complemented by methods for affinity prediction, which aim at estimation of either absolute or relative binding free energies, usually using more sophisticated biophysical models. These methods are briefly reviewed in section 2.4.2. The FlexScreen approach developed in our group (Merlitz and Wenzel 2004; Fischer, Merlitz et al. 2005; Wenzel, Fischer et al. 2007; Wenzel, Fischer et al. 2008), which interpolates between these two families by using biophysical forcefields and partial receptor flexibility in a docking context, is discussed in section 5.2 of the applications chapter.

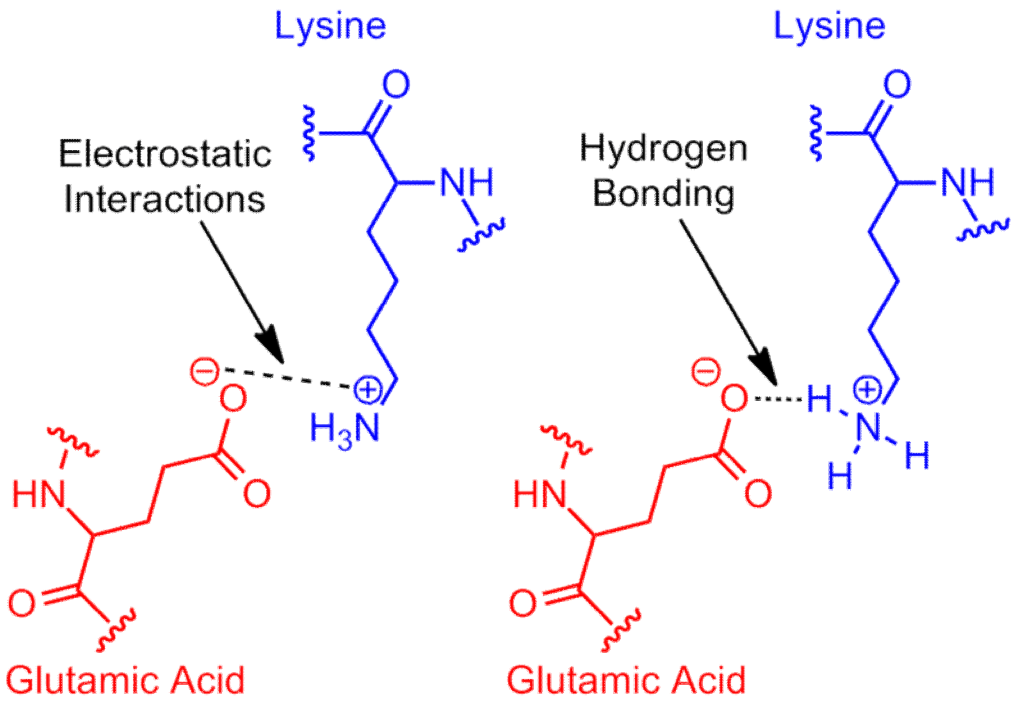


Figure 2.9: Example of salt bridge between amino acids glutamic acid and lysine demonstrating electrostatic interaction and hydrogen bonding, (taken from http://en.wikipedia.org/wiki/Salt_bridge_%28protein_and_supramolecular%29).

2.4.1. Docking Methods

Docking is a computational method which predicts the preferred orientation of one molecule to a second (Lengauer and Rarey 1996). Such methods have been applied in a number of contexts: e.g. protein-protein, protein-DNA and protein-ligand docking. Here we focus on protein-ligand docking, where a small molecule binds to a protein receptor.

In the past, two complementary approaches have been investigated: one, where the protein and the ligand are describe as complementary surfaces (Jorgensen 1991; Kitchen, Decornez et al. 2004); while the second approach evaluates an approximation of the binding energy (scoring function) for every conformation of the complex (Wei, Weaver et al. 2004). In the second approach, which we discuss in the following, the protein and the ligand conformational spaces are explored with a search algorithm to optimize the scoring function. Typical moves in such algorithms include translations and rotations (rigid body transformations), and torsion angle rotations (internal changes of the ligand). Finally the scoring function for the predicted complex is calculated, yielding an estimate of the affinity.

This method has some advantages, as for example that the process is very similar to the docking process in reality, but the scoring functions used today often give poor approximations of the affinity (Warren, Andrews et al. 2004). Depending on the docking speed the method can be employed for screening large compound databases in the search for drug compounds. One major complication in docking methods are treatment of the conformational flexibility of the receptor and treatment of induced fit effects (Matthews, Wei et al. 2004; Wenzel and Kokh 2008).

Generally docking methods involve generation of a set of poses for a ligand that can fit the binding site. The poses are ranked based on a scoring function (Mobley and Dill 2009). There are different scoring functions that can be used depending on the desired approach. There are regression-based or interaction-based scoring functions. Interaction-based scoring functions are similar to molecular mechanics forcefields, for example the CHARMm (Brooks, Brooks et al. 2009) and AMBER (Cornell, Cieplak et al. 1995), which have been used to calculate the enthalpy of binding. Normally the values of non-bonded energy terms (van der Waals, electrostatic and internal energy related to bond angles, bond lengths, torsional angles) are precalculated on a grid which is used to compute the energy contributions for atoms placed in the receptor pocket. Moreover an approximation of solvent effects can be added. Using such scoring functions or forcefields the ligand deformation is treated in the same way as the interaction between the ligand to the protein.

Knowledge-based potentials improve modeling of protein complexes by taking advantage of the rapidly increasing amount of experimentally derived information on protein-protein association. An essential element of knowledge-based potentials is defining the reference state for the optimal description of residue-residue pairs in the non-interaction state.

Studies in scoring functions showed that XSCORE (Obiol-Pardo and Rubio-Martinez 2007) performs better than the other empirical scoring functions (Wang, Lu et al. 2003) and is one of the most used programs to assess ligand-binding affinity. This program is a consensus scoring function, resulted from the arithmetical average of three empirical scoring functions.

In addition, the use of the cross-term empirical scoring function implemented in the program POLSCORE (Dias, Timmers et al. 2008) was able to predict the orientation of ligands for complexed structures with results better than XSCORE (Obiol-Pardo and Rubio-Martinez 2007; Kulharia, Goody et al. 2008) and DrugScore (Dias, Timmers et al. 2008).

Many efforts have been made to develop techniques to speed up the search for the best conformation. For example, Monte-Carlo methods (Goodsell, Huey et al. 2007) or its generalizations (Wenzel and Hamacher 1999) are often used, while other methods use genetic algorithms, as for example in the program Gold (Jones, Willett et al. 1997). Fragment based approaches use a different approach to sample the large conformational space of a protein-ligand complex. Different fragments of the ligand are created and then independently docked to the protein to be finally assembled again (Taylor, Jewsbury et al. 2002).

Glide uses a series of hierarchical filters to search for possible locations of the ligand in the active-site region of the receptor (Friesner, Banks et al. 2004).

For the reconnection of the broken bonds the most popular approach is the incremental construction algorithm implemented in FlexX (Rarey, Kramer et al. 1996). The *de novo* ligand design methodology is linked to this approach where a totally new ligand is constructed by docking fragments of a database to the protein instead of screening a database of ligands (Sandor, Kiss et al. 2010).

2.4.2. Estimates of the Affinity

Evaluate the affinity of protein-ligand interactions with reliable methodologies has become very important view the big impact of drug discovery to detect new lead compounds, and chemical genomics to search for inhibitors to elucidate gene function (Gilson and Zhou 2007).

To increase the impact for these computational methodologies one needs experimental determined affinities to correlate with the predicted affinities. The experimental determined affinities are used as guidelines to calibrate these empirical scoring functions. Free energy studies in the last years have been conducted using relative or absolute binding free energy calculations (Guimaraes, Boger et al. 2005), explicit or implicit solvent models (Fujitani, Tanida et al. 2005; Michel, Verdonk et al. 2006), etc…

A challenging task is to calculate the absolute binding free-energies. There are different options available such as the free energy perturbation (FEP) (Zwanzig 1954). This approach is at least formally rigorous and capable of considering relaxation process in the protein/solvent system. On the other hand this method often remains computationally prohibitive because the FEP require calculations on large macromolecular assemblies surrounded by explicit water molecules. Moreover, the available computer power does not allow one to use the FEP approaches in effective docking studies where different binding sites should be explored.

An alternative to the FEP calculations can be obtained by using the microscopic all-atom linear response approximation (LRA) (Lee, Chu et al. 1992) for the electrostatic part of the thermodynamics cycle and approximating the non-electrostatic part by different strategies. Another related approach is the widely used linear interaction energy (LIE) method (Hansson, Marelius et al. 1998), which adopts the LRA approximation for the electrostatic contribution while estimating the non-electrostatic term by scaling the average van der Waals (vdW) interaction. A significantly faster option is offered by the semi-macroscopic protein dipoles langevin dipoles (PDLD/S) in its LRA form (PDLD/S-LRA) (Sham, Chu et al. 2000). Another commonly used semi-macroscopic method is the molecular mechanics Poisson-Boltzmann/surface area (MM/PBSA) (Cheatham, Srinivasan et al. 1998). This approach combines molecular mechanics (MM) energy, solvation free energies with Poisson Boltzmann calculations, and entropy estimates from quasiharmonic (QH) or normal mode (NMODE) analysis to predict the absolute binding free energy. Because of its unclear theoretical foundation this method has drawn some criticism (Pearlman 2005). MM/PBSA and the MM/GBSA methods are an apparent adaptation of the PDLD/S-LRA idea of MD generation of conformations for implicit solvent calculations, but these methods only calculate the average over the configurations generated with the charged solute while ignoring the uncharged term. In the MM/PBSA approach, attention has been focused to adjusting model parameters, such as atomic radii and solute dielectric constant, to reproduce experimental observations, rather than looking for more physical implicit solvent representations.

Accurate estimation of the free energy changes still remains a challenge despite the recent developments in computing power and methodology. The level of description included in different computational models allows a compromise between the simplicity, accuracy and the performance of the calculations. The main concern that remains is how to accommodate the number of degrees of freedom to accurately describe the protein/ligand system.

2.5. Protein Protein Interactions

Similar to methods studying protein-ligand interactions for small molecule ligands, much work has been invested to develop models to predict conformation and affinity of protein-protein complexes (Stoddard and Koshland 1992; Smith and Sternberg 2002; Wiehe, Peterson et al. 2008), which are important regulators of biological function. Such methods are either based on empirical bioinformatics approaches, such as functional matrices, (Marrero-Ponce, Medina-Marrero et al. 2005) or on molecular modeling techniques, such as molecular mechanics MM Generalized Born/Surface Area GBSA (Massova and Kollman 1999), similar to those discussed already in the previous section.

Protein-protein interactions are essential to understanding the function of biological systems, and their characterization has become an important task for both experimental and computational approaches in systems biology. Experimental methods include yeast two-hybrid systems (Uetz, Giot et al. 2000; Ito, Chiba et al. 2001), mass spectrometry (Ewing, Chu et al. 2007) and protein chips.

Computational methods are based on simple sequence and genomic features intuitively related to interactions and functional relationships. In the following I will focus on protein-protein docking and alanine scanning mutagenesis, which are of direct relevance to the work reported in this thesis.

2.5.1. Protein Protein docking

Protein docking is generally applied to individual pairs of proteins that are known to interact (Fig. 2.10). Two principal issues must be addressed to solve this problem; a scoring function/energy function must be developed that can discriminate correctly or near-correctly docked orientations from incorrectly docked ones, and a search method must be implemented to ‘find’ a near-correctly docked orientation. The net problem is much more complicated than small-molecule ligand docking simulations, because in protein-protein docking the docking site is generally not known. Therefore all respective orientations of the molecules with respect to each other must be sampled, which is difficult with a standard dynamic and methods. Instead methods exploiting surface shape complementarity (simplest scoring function) are used as an initial step, which is defined later. This may be done by discretising the molecule into a grid in space and considering which cells are occupied, or by using some sort of ‘surfacing algorithm’, which calculates the solvent-accessible or solvent-excluded surface, and a point set that triangulates.

Most docking programs comprise two standard steps: generation of thousands of alternative poses to sample all possible interaction modes, followed by scoring these poses using an energy function. In many cased conformations very similar to the natural one are generated by the first step, but scoring functions often fail to rank them properly (Lensink, Mendez et al. 2007). The fast Fourier transform (FFT) method is the most used technique for the first stage of docking. This approach is used in the programs GRAMM (Vakser 1995), FTDock (Gabb, Jackson et al. 1997), ZDOCK, 3D-Dock (Sternberg, Aloy et al. 1998) and DOT (Mandell, Roberts et al. 2001) (Table 2.II).

The rigid-body approximation is abandoned after the two first steps to introduce flexibility. Flexibility is introduced, at least in the protein sidechains and possibly also in the backbone. In most methods a molecular mechanics forcefield is used to minimize the energy of the complex. This part of the simulation is very similar to the simulations used for small-molecule docking, because here the overall shape of the complex is already known and changes little during the simulation. In the following I have compiled a brief overview of some of the methods most used in this field.

Program	Algorithm	URL
3D-Dock	Global: FFT; rescoring: residue potentials; refinement: mean-field sidechain multicopy	www.bmm.icnet.uk/ docking/
HEX	Global: Fourier correlation of spherical harmonics	www.biochem.abdn.ac.uk/hex/
GRAMM	Global: FFT clustering and rescoring	reco3.ams.sunysb.edu/gramm/
DOT	Global: FFT for shape complementarity and approximate Poisson-Boltzmann electrostatics	www.sdsc.edu/CCMS/DOT
BIGGER (chemera)	Global: bit mapping; rescoring: multiple filters	www.dq.fct.unl.pt/bioin/chemera/
DOCK	Global: grid-based energy function; flexible docking: random search plus incremental construction	www.cmpharm.ucsf.edu/ kuntz/dock.html
AutoDock	Grid-based empirical potential Scripps Institute flexible docking via Monte Carlo search and incremental construction	http://mgl.scripps.edu/
FlexX	Fragment assembly energy function: (Boehm potential)	cartan.gmd.de/flexx/

Table 2.II: Programs for protein-protein docking.

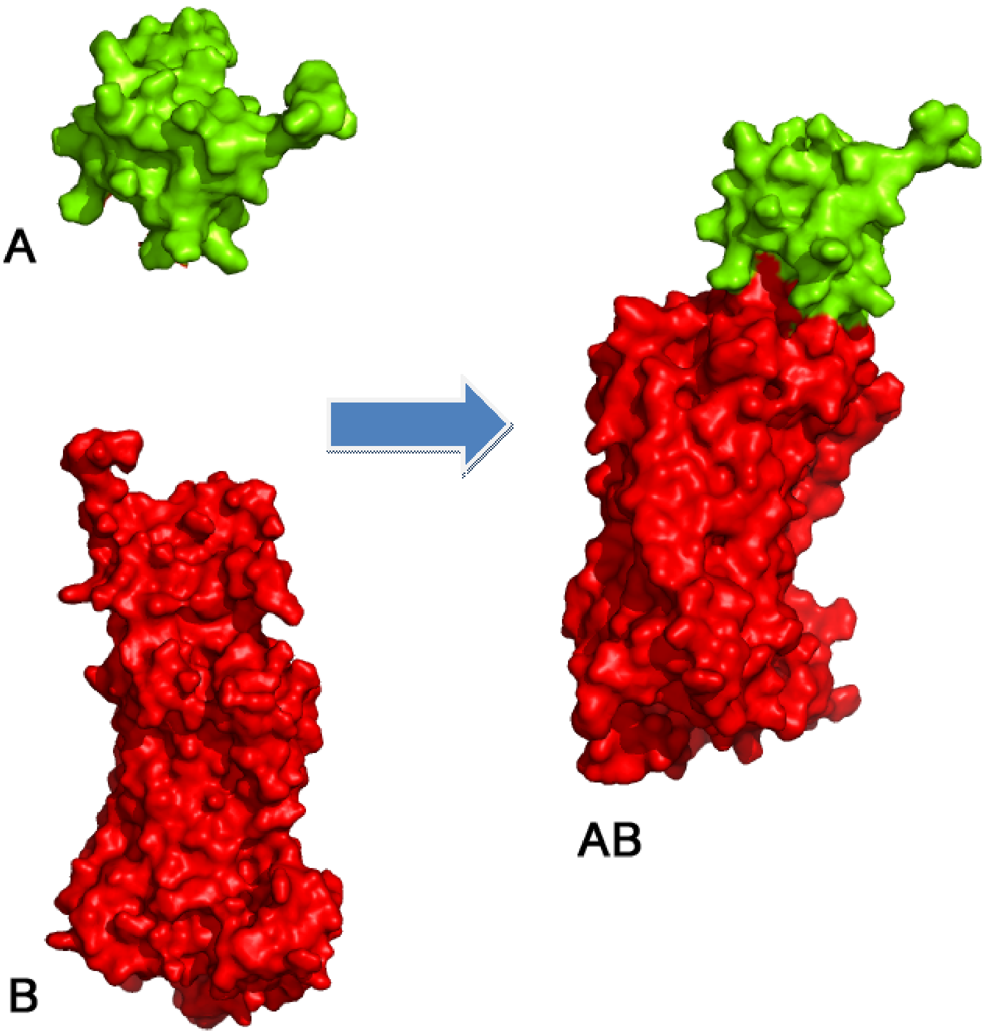


Figure 2.10: example of docking, (A) and (B) (Receptor-Ligand) are 2 separated structures. (AB), Docking predict preferred orientation of one molecule to another.

2.5.2. Alanine scanning mutagenesis

Knowledge about affinity and specificity in protein interfaces can be used to inhibit protein interactions. Experimental alanine scanning mutagenesis is a powerful tool for analyzing interactions in protein interfaces (De Genst, Areskoug et al. 2002) and helps to identify hotspots of protein-protein interactions. Scanning all amino acids of a protein-protein interface can allows to create a map of which interactions are critical for protein binding and which ones are not. As well known, protein-protein complex formation depends, in most cases, on only a few interface residues, called ‘hotspots’ (Bogan and Thorn 1998) that account for the highest contribution to the binding free energy (Clackson and Wells 1995; Ofran and Rost 2007). Alanine scanning still represents a very important experimental effort, even though it cannot be applied easily to high-throughput screening of protein-protein interfaces.

For this reason also computational alanine scanning methods have been developed, which calculate the change in binding free energy ($\Delta\Delta G$) of a protein-protein complex after mutation of an amino acid residue with alanine. Computational prediction methods are used to complement experiments and to enhance the understanding of protein stability. Alanine scanning methods such as Robetta (Kortemme and Baker 2002), FoldEF (Guerois, Nielsen et al. 2002) and knowledge-based methods (K-FADE, K-CON) (Darnell, Page et al. 2007; Darnell, LeGault et al. 2008) have been proposed.

Thermodynamic simulations are mostly used to estimate the free energy of association. Although these methods include energy terms, which are important for protein stability, there is still a large discrepancy between predicted values and experimentally measured free energy changes. Recently, a knowledge-based model was introduced to predict binding ‘hot spots’ (Darnell, Page et al. 2007; Darnell, LeGault et al. 2008), but the prediction accuracy was relatively low. The Rosetta fragment-based approach has also been extended to study protein-protein interactions based on sequence information and homology to proteins of known structure combined with a *de novo* protocol for non-homologous portions of the protein. Rosetta creates protein structures as well as energies for wild-type proteins and alanine screens of protein complexes with available sequence information (Simons, Kooperberg et al. 1997; Chivian, Kim et al. 2003). The empirical algorithm FoldEF carries out free-energy calculations based on three-dimensional structural data using experimental terms for interactions and weighting factors for energy terms derived from experimental data sets. It can be used for computational alanine screening (CAS) by comparing interaction energies of wild-type and mutated complexes (Guerois, Nielsen et al. 2002; Schymkowitz, Borg et al. 2005).

In the MM-GBSA approach interaction free energies of individual residues in a protein-protein complex are estimated by calculating gas-phase energies, solvation free energies, and entropic contributions for the free proteins and the complex derived from selected snap shots of the trajectories (Huo, Massova et al. 2002; Simonson, Archontis et al. 2002; Gohlke, Kiel et al. 2003; Benedix, Becker et al. 2009). Molecular dynamics simulations have also been used to detect and characterize hot spots as conserved residues have generally restricted flexibility in the unbound state (Rajamani, Thiel et al. 2004; Yogurtcu, Erdemli et al. 2008).

Many efforts have been made to identify correlations between binding hot spots and protein structure and sequence information (Hu, Ma et al. 2000; Ma, Elkayam et al. 2003; Halperin, Wolfson et al. 2004). These methods assume that structurally conserved residues are strongly correlated with interaction hot spots and that hot spots are not compactly clustered but distributed over the interface. Moreover, the identification of similar residues in hot spots in various protein families may suggest that affinity and specificity are not necessarily coupled

(Ma, Elkayam et al. 2003). Several studies have examined hot spots where systematic analysis of the structural features is limited to the solvent accessibility and surface area between the unbound and bound states (ΔASA). Qualitative analyses have been performed, but not statistical analysis has yet been applied (Cho, Kim et al. 2009). The qualitative nature of the analyses performed to date mainly derives from the difficulty of identifying features that distinguish hot spots from other residues in interaction interfaces.

3. Application of protein structure prediction to identify genetic causes for human disease

3.1. Introduction

The Human Genome Project (HGP) was a collaborative research program between international groups. The human genome project formally begun in October 1990 and completed in 2003, it succeeded to sequence all the estimated 20,000-25,000 human genes and make them accessible for further biological study (Luria 1989; Collins 1997; Rowen, Mahairas et al. 1997). The complete genome is a resource of detailed information about the structure, organization and function of the complete set of human genes. This information can be thought of as the basic set of inheritable "instructions" for the development and function of a human being. Sequencing the human genome raised high hopes to identify causes for many diseases and developmental disorders at the genetic level. Ten years after the end of the human genome project, these hopes are much more subdued (Barnhart 1988; Roach, Boysen et al. 1995; Venter 2010; Venter 2011). However, there have been notable exceptions, such as regulatory sequences, non-coding DNA (Carroll, Prud'homme et al. 2008) and the mapping of a personal genome (Roukos 2009), motivating a continued search for single gene defects that may be involved in specific diseases. Because the cost of sequencing has been reduced significantly, it is now possible to sequence significant fractions of the genome of individual patients that are afflicted by a particular disease.

In this chapter I report on a joint experimental/theoretical, effort to investigate possible causes for one human developmental disorder, Idiopathic hypogonadotropic hypogonadism (IHH) or Kallmann syndrome (KS). Idiopathic hypogonadotropic hypogonadism (IHH), or normosmic IHH (nHH) constitutes one of the most common forms of congenital hypogonadism (Layman, Cohen et al. 1998). Humans with nHH suffer from absent pubertal development due to impaired GnRH secretion and/or action, low serum testosterone (males) or estradiol (females), and low FSH and LH without other pituitary pathology (Dode, Levilliers et al. 2003; Kim, Herrick et al. 2005; Pitteloud, Acierno et al. 2006). Kallmann syndrome (KS) consists of the combination of nHH with anosmia due to impaired migration of GnRH and olfactory neurons. Associated anomalies may include synkinesia (mirror movements), ataxia, visual abnormalities, hearing loss, mental retardation, dental agenesis, midfacial defects, and renal agenesis (de Roux, Young et al. 1997; Kim, Bhagavath et al. 2008). Inheritance patterns may be autosomal dominant, autosomal recessive, and X-linked recessive, or it may be sporadic. In addition, several cases of digenic disease have been reported (Layman, Cohen et al. 1998; Seminara, Messager et al. 2003).

Human genetic studies of nHH and KS have yielded important insights into the identification of causative genes. At least one mutation has been reported in a number of human genes. In this chapter I report work on three genes possibly involved in the disorder, which encode the proteins *CHD7*, *WDR11* and *NELF*. Given the genetic information on those three proteins, I have investigated the observed mutations with respect to their tendency to be detrimental to protein function using bioinformatics based methods and protein structure prediction to support the experimental work. All work reported in this chapter was performed in a collaboration involving groups at: Molecular Neurogenetics Unit, Center for Human Genetic Research, Massachusetts General Hospital; Department of Genetics, Harvard Medical School; Institut für Humangenetik, Universitätsklinikum Hamburg; Department of Pathology, Harvard Medical School; Department of Life Science, Cell Dynamics Research Center, Gwangju, South Korea; Division of Clinical Molecular Pathology and Genetics, Department of Pediatrics, Ankara; The Medical College of Georgia; Neuroscience Neurobiology Programs in the Institute of Molecular Medicine and Genetics; and Neuroscience Program, Medical College of Georgia; Department of Human Molecular Genetics, Max Planck Institute for Molecular Genetics,Berlin; Department of Biology and Graduate Division of Basic Medical Sciences, St. George's Medical School, University of London; Division of Endocrinology, Diabetes and Metabolism, Cedars-Sinai Medical Center, Los Angeles; Reproductive Endocrinology and Infertility, Women and Infants Hospital, Medical College of Georgia, coordinated by the experimental groups of L. Layman (U. Georgia) and J. Gusella (Massachusetts General Hospital/Harvard Medical School) and has been reported in part in the following publications: (Kim, Kurth et al. 2008; Kim, Ahn et al. 2010; Xu, Kim et al. 2011). I have performed all the modeling work for these studies, but report also key experimental findings in the subsequent sections to put these into the perspective of the overall study.

In the following I will first summarize the methods I have used (section 3.2) and then devote one section each to *CHD7* (section 3.3), *WDR11* (section 3.4), *NELF* (section 3.5). In section

3.6 I summarize all findings and provide an assessment of the state-of-the-art and an outlook for further work.

3.2. Methods

An overview of the general principles of computational methods for protein structure prediction has been given in section 2.3.1. To support the experimental work reported in this chapter I have applied methods of homology based protein structure prediction, the specific protocols were similar for all studies in this chapter and are summarized here. For each target, templates were selected using PHYRE (Kelley and Sternberg 2009), FUGUE (Shi, Blundell et al. 2001) and 3DJURY (Ginalski, Elofsson et al. 2003). The structural model of the sequence is built based on the alignment with the chosen template using ClustalW (Higgins, Thompson et al. 1996) and EXPRESSO (Armougom, Moretti et al. 2006). ClustalW calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen (Higgins and Sharp 1988; Thompson, Higgins et al. 1994; Aiyar 2000). The methods are based on protein threading, which exploits the fact that the number of different folds observed in nature is quite small (1000); and that almost all the new structures submitted to the PDB in the past three years are similar in structural folding to the ones already in the PDB (Wang, Fang et al. 2004; Berman, Henrick et al. 2007). I then used MOE (Vilar, Cozza et al. 2008) to generate a model on the basis of the alignment.

In a second set of analysis, we investigated whether specific mutations have a tendency to be detrimental to protein function using bioinformatics based methods, specifically we used POLYPHEN (Flanagan, Patch et al. 2010), SIFT (Ng and Henikoff 2003). In addition we have used SPPIDER (Porollo and Meller 2007) to identify possible binding sites in the sequence. SIFT (Sorting Intolerant From Tolerant) predicts whether substitution of an amino acid can affects protein function based on sequence homology and on the physical properties of amino acids. SIFT can be used to non-synonymous polymorphisms and to missense mutations tested in laboratory. The program presumes that important amino acids will be conserved in the protein family, and so changes at well-conserved positions tend to be predicted as deleterious. For example, if a position in an alignment of a protein family only contains the amino acid isoleucine, it is presumed that substitution to any other amino acid is selected against and that isoleucine is necessary for protein function. Therefore, a change to any other amino acid will be predicted to be deleterious to protein function. The SPPIDER protein method consider the single chain of protein resolved 3-D structure to predict the residues that could be find at the protein interface. SPPIDER is able to determine the residues that are potentially involved in protein-protein interactions (based on consensus classifier) where the predicted interacting sites may belong to the different protein interfaces.

The POLYPHEN (=Polymorphism Phenotyping) algorithm, uses a set of empirical rules based on sequence, phylogenetic, and structural information predicts how damaging a particular variant could be. POLYPHEN utilizes not only sequence alignment, but also protein structure databases, such as PDB (Protein Data Bank) or PQS (Protein Quaternary Structure), DSSP (Dictionary of Secondary Structure in Proteins), and three-dimensional structure databases to determine if a variant may have an effect on the protein's secondary structure, interchain contacts, functional sites, and binding sites. POLYPHEN assign as "probably damaging", a high confidence score where the nsSNP should affect protein structure and/or function; "possibly damaging", where it may affect protein function and/or structure; and "benign", as most likely having no phenotypic effect. The determination of the score is due to the position-specific independent count difference of the two allelic variants in the polymorphic position and to the degree of damaging effect a variant may have on structural parameters of a protein. The position-specific independent count is a logarithmic ratio of the likelihood of a given amino acid occurring at a particular position to the likelihood of the same amino acid occurring at any position in the sequence.

3.3. Investigation of Mutations in *CHD7*

Mutations in the chromodomain helicase DNA-binding protein 7 (*CHD7*) gene have been identified in patients with CHARGE syndrome, a multisystem, autosomal dominant disorder (Vissers, van Ravenswaaij et al. 2004). Our experimental colleagues identified seven heterozygous *CHD7* mutations, two splice and five missense mutations, in three sporadic KS and four sporadic normosmic IHH patients. Three mutations affect the chromodomain, critical for proper *CHD7* function in chromatin remodeling and transcriptional regulation of genes, suggesting they are deleterious and causative. *CHD7*'s role is further corroborated by specific expression in IHH/KS-relevant tissues, gonadotropin releasing hormone (GnRH) cell lines, and appropriate developmental expression. *CHD7* represents the first chromatin remodeling protein found to have a role in human puberty and the second gene identified to cause both normosmic IHH and KS in humans, but no structure for the protein is available.

In mammals, chromodomain proteins appear to be either structural components of large macromolecular chromatin complexes or proteins involved in remodeling chromatin structure. Recent work has suggested that apart from a role in regulating gene activity, chromodomain proteins may also play roles in genome organization (Jones, Cowell et al. 2000).

Transcriptional regulation in an eukaryotic nucleus requires cellular activities that recognize and act on chromatin (Orlando and Jones 2002; Steger, Haswell et al. 2003), hence chromodomain 1 could be critical for protein function.

The *CHD7* structure with functional domains and the positions of five missense mutations and two splice-donor site mutations identified in IHH and KS patients are shown in figure 3.1. The function of CR1-CR3 and BRK domains are unknown. Three of the observed mutations affect chromodomains.

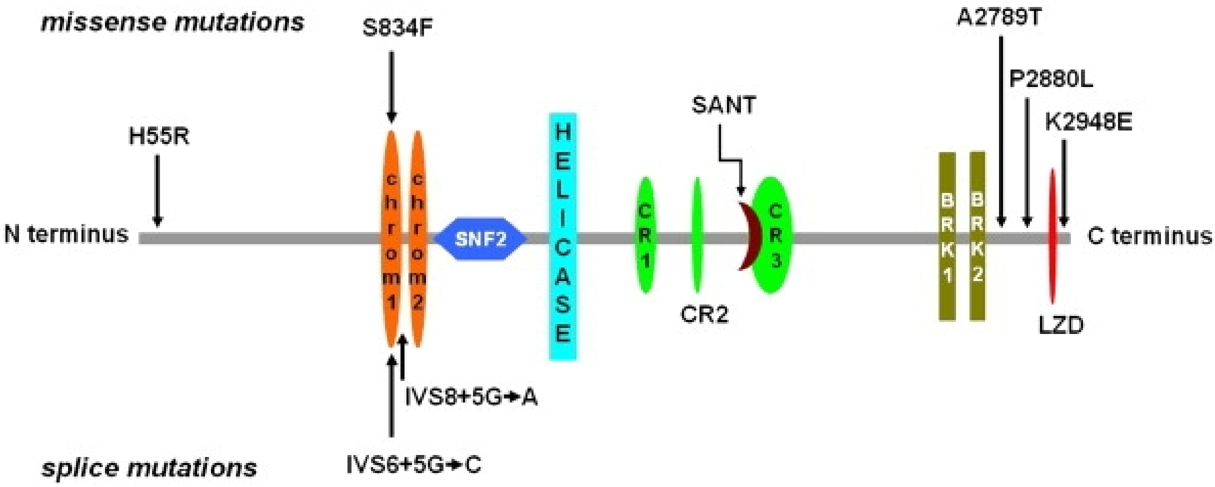


Figure 3.1: *CHD7* Domains and Positions of Mutations. The following abbreviations are used: chrom1, chromodomain 1; chrom2, chromodomain 2; SANT, SANT DNA binding domain; and LZD, leucine zipper domain. Relative sizes and locations of domains are to scale.

3.3.1. Sequence Based Analysis of Mutations

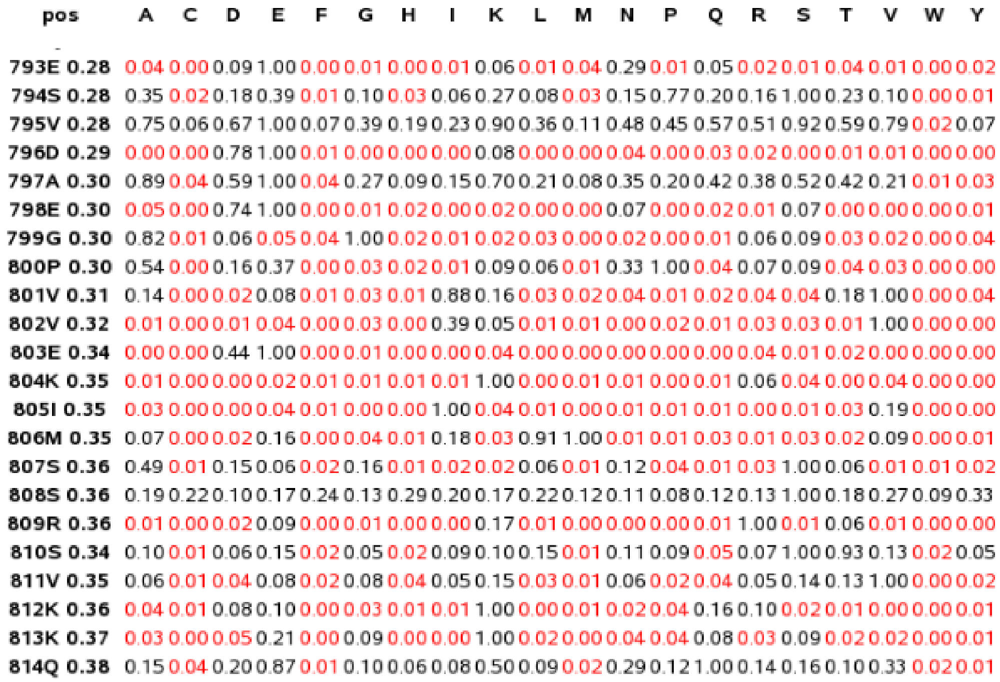
We performed a SIFT analysis for *CHD7* to investigate potentially deleterious effects of point mutations, in frame deletion and a truncation mutations. The SIFT approach uses homology to predict whether an amino acid substitution will affect protein function and hence, potentially alter phenotype. Among the five point mutations investigated (Fig. 3.1) four were predicted to be deleterious (Ser834Phe, Lys2948Glu, Pro2880Leu, His55Arg), while not effect was predicted for Ala2789Thr. The prediction of the effect of Ser834Phe, located in chromodomain 1 DNA-binding (UniProtKB/Swiss-Prot entry Q9P2D1) (Wu, Apweiler et al. 2006), had the highest confidence measure (see methods). The chromodomain 1 (INTERPRO IPR000953 (Mulder, Apweiler et al. 2003; Mulder and Apweiler 2008)) is a highly conserved sequence motif that has been identified in a variety of animal and plant species. Analyses of structural models for the relevant domains suggest that Ala2789Thr and Pro2880Leu, located in the spacer sequence between the BRK2 and leucine-zipper regions, as well as Lys2948Glu, are also detrimental.

SIFT was also used to characterize the functional significance of a 22 amino acid (residues: ESVDAEGPVVEKIMSSRSVKKQ) in-frame deletion in exon 6 (from 793 to position 814), by predicting deleterious effects for all possible amino acid substitutions in this region (Table 3.I). Deleterious effects were predicted for essentially all point mutations in this region, thus deleterious effects will be observed with high likelihood for the deletion of the whole domain. The truncation mutation (starting from amino acid 810 with the following domains deleted) eliminates the region from amino acid 920-1490, which are highly homologous to the chromatin remodeling domain in SWI2/SNF2 chromatin remodeling domain of eukaryotic Rad54 (PDB code: 1Z3I) (Thoma, Czyzewski et al. 2005) and thus the essential DNA-binding portion of the protein.

3.3.2. Structural Models of *CHD7* Domains

For most of the *CHD7* protein, sequence homology is too low to generate a reliable high-resolution model. However, combining different techniques we were able to generate a local structural models for the 155 AA region around Ser834Phe (795-950) and for the last 300 amino acids of the C-terminus of *CHD7* starting after the BRK domain and including the leucine finger. Alignments were generated by PHYRE, FUGUE and 3DJURY. For the first region a perfect homology (>99% confidence by PHYRE) to the yeast CHD1 tandem chromodomains (PDB code 2h1eA (Flanagan, Blus et al. 2007)) was found for the relevant region and converted into a 3-D model (Fig. 3.2). Regions of the model that were predicted solely on the basis of homology in 3DJURY are shown in green, but additional information was used to substantiate the model.

For the second region no single approach yielded sufficient homology to build a model, but 3DJURY found an alignment with the crystal structure of mdia1 gbd-fh3 in complex with rhoc (1z2c/B) a signaling protein. The sequence identity of this alignment is insufficient to validate the model, but with EMBOSS we found an alignment with 19% identical and 35% similar residues (colored sky blue in the model). To further validate the model we performed a secondary structure prediction using PHYRE, which computes a consensus prediction based on psipred, jnet, sspro. The independently predicted helical regions (colored in wheat) agree well with model, as does the leucine zipper domain (colored in blue). The mutations Lys2948Glu, Pro2880Leu and Ala2789Thr are all located in highly disordered loop regions most likely to affect structural and binding properties of the domains to other *CHD7* domains or interaction partners. For the last 300 AA no single approach yielded sufficient homology to build a model, so we used different approaches using again 3DJURY, PHYRE and also EMBOSS.



pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
793E 0.28	0.04	0.00	0.09	1.00	0.00	0.01	0.00	0.01	0.06	0.01	0.04	0.29	0.01	0.05	0.02	0.01	0.04	0.01	0.00	0.02
794S 0.28	0.35	0.02	0.18	0.39	0.01	0.10	0.03	0.06	0.27	0.08	0.03	0.15	0.77	0.20	0.16	1.00	0.23	0.10	0.00	0.01
795V 0.28	0.75	0.06	0.67	1.00	0.07	0.39	0.19	0.23	0.90	0.36	0.11	0.48	0.45	0.57	0.51	0.92	0.59	0.79	0.02	0.07
796D 0.29	0.00	0.00	0.78	1.00	0.01	0.00	0.00	0.00	0.08	0.00	0.00	0.04	0.00	0.03	0.02	0.00	0.01	0.01	0.00	0.00
797A 0.30	0.89	0.04	0.59	1.00	0.04	0.27	0.09	0.15	0.70	0.21	0.08	0.35	0.20	0.42	0.38	0.52	0.42	0.21	0.01	0.03
798E 0.30	0.05	0.00	0.74	1.00	0.00	0.01	0.02	0.00	0.02	0.00	0.00	0.07	0.00	0.02	0.01	0.07	0.00	0.00	0.00	0.01
799G 0.30	0.82	0.01	0.06	0.05	0.04	1.00	0.02	0.01	0.02	0.03	0.00	0.02	0.00	0.01	0.06	0.09	0.03	0.02	0.00	0.04
800P 0.30	0.54	0.00	0.16	0.37	0.00	0.03	0.02	0.01	0.09	0.06	0.01	0.33	1.00	0.04	0.07	0.09	0.04	0.03	0.00	0.00
801V 0.31	0.14	0.00	0.02	0.08	0.01	0.03	0.01	0.88	0.16	0.03	0.02	0.04	0.01	0.02	0.04	0.04	0.18	1.00	0.00	0.04
802V 0.32	0.01	0.00	0.01	0.04	0.00	0.03	0.00	0.39	0.05	0.01	0.01	0.00	0.02	0.01	0.03	0.03	0.01	1.00	0.00	0.00
803E 0.34	0.00	0.00	0.44	1.00	0.00	0.01	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.04	0.01	0.02	0.00	0.00	0.00
804K 0.35	0.01	0.00	0.00	0.02	0.01	0.01	0.01	0.01	1.00	0.00	0.01	0.01	0.00	0.01	0.06	0.04	0.00	0.04	0.00	0.00
805I 0.35	0.03	0.00	0.00	0.04	0.01	0.00	0.00	1.00	0.04	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.03	0.19	0.00	0.00
806M 0.35	0.07	0.00	0.02	0.16	0.00	0.04	0.01	0.18	0.03	0.91	1.00	0.01	0.01	0.03	0.01	0.03	0.02	0.09	0.00	0.01
807S 0.36	0.49	0.01	0.15	0.06	0.02	0.16	0.01	0.02	0.02	0.06	0.01	0.12	0.04	0.01	0.03	1.00	0.06	0.01	0.01	0.02
808S 0.36	0.19	0.22	0.10	0.17	0.24	0.13	0.29	0.20	0.17	0.22	0.12	0.11	0.08	0.12	0.13	1.00	0.18	0.27	0.09	0.33
809R 0.36	0.01	0.00	0.02	0.09	0.00	0.01	0.00	0.00	0.17	0.01	0.00	0.00	0.00	0.01	1.00	0.01	0.06	0.01	0.00	0.00
810S 0.34	0.10	0.01	0.06	0.15	0.02	0.05	0.02	0.09	0.10	0.15	0.01	0.11	0.09	0.05	0.07	1.00	0.93	0.13	0.02	0.05
811V 0.35	0.06	0.01	0.04	0.08	0.02	0.08	0.04	0.05	0.15	0.03	0.01	0.06	0.02	0.04	0.05	0.14	0.13	1.00	0.00	0.02
812K 0.36	0.04	0.01	0.08	0.10	0.00	0.03	0.01	0.01	1.00	0.00	0.01	0.02	0.04	0.16	0.10	0.02	0.01	0.00	0.00	0.01
813K 0.37	0.03	0.00	0.05	0.21	0.00	0.09	0.00	0.00	1.00	0.02	0.00	0.04	0.04	0.08	0.03	0.09	0.02	0.02	0.00	0.01
814Q 0.38	0.15	0.04	0.20	0.87	0.01	0.10	0.06	0.08	0.50	0.09	0.02	0.29	0.12	1.00	0.14	0.16	0.10	0.33	0.02	0.01

Table 3.I: SIFT scores of the tolerated single point mutations for the in-frame deletion.

Looking at the mutations, we found that all three AA residues are located in loop regions so that mutations of these residues will most likely affect structural and binding properties of the domains to their interaction partners (Fig. 3.2 A and 3.2 B). The local secondary structure of the region around Pro2880 is a random coil, and the Pro2880Leu mutation induces helix formation in this region, predicting a deleterious effect (data not shown). The Ser834 residue is also located in a loop region, coordinating strongly with adjacent residues in the neighboring helix (Tyr881, context PDYV), so Phe834 may therefore strongly affect protein stability.

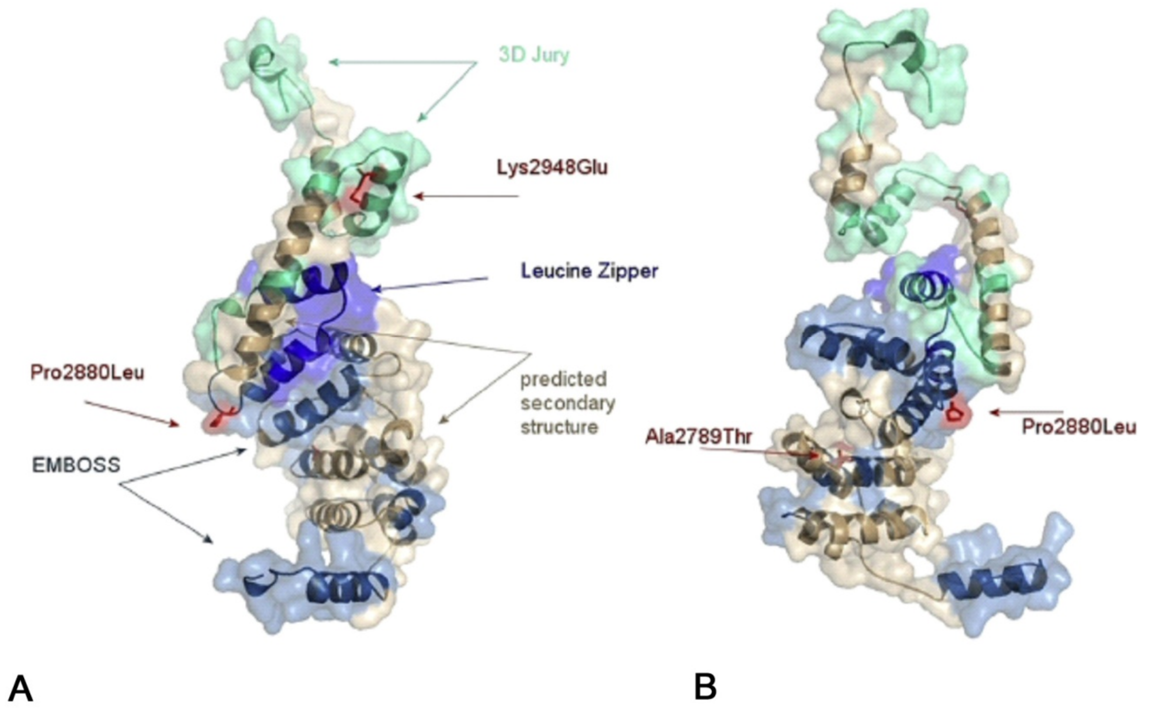


Figure 3.2: ***CHD7*** **structure modeling.**
(A and B) Shown are alternate views of the model of the 300 AA C-terminal region of ***CHD7*** **based on the 3DJURY model, which results from the alignment of this region with that of mdia1 gbd-fh3 in complex with rhoc (1z2c/B). (A) and (B) show the same model rotated 180 around the vertical axis though the center of the molecule.**

3.3.3. Conclusions

The theoretical work performed for *CHD7* is a supportive evidence that the *CHD7* experimental mutations are deleterious as predicted from SIFT analysis, which indicates that four of the five missense mutations (Ser834Phe, Lys2948Glu, Pro2880Leu, and His55Arg) involve highly conserved AA residues among known species and, therefore, are not likely to be tolerated by their observed substitutions (Ng and Henikoff 2003). These findings were also corroborated by protein structural analysis of the AA variants Ala2789Thr, Pro2880Leu, and Lys2948Glu, which were predicted to alter structural and binding properties of the domains. Taken together, both AA conservation and protein structural modeling provide additional support that these missense substitutions are highly likely to be deleterious mutations.

Importantly, one IHH and one KS patient, who both lack the CHARGE phenotype, possess the same mutations (Ser834Phe and IVS6þ5G/C) reported in patients with CHARGE syndrome (Delahaye, Sznajer et al. 2007; Jongmans, Hoefsloot et al. 2008), further demonstrating the allelic relationship of both syndromes. The KS patient with the IVS6þ5G/C mutation does not fulfill Blake's criteria for CHARGE syndrome (Blake, Salem-Hartshorne et al. 2005), although she does have hearing impairment and cleft lip and palate.

This also indicates that the effects of modifying genes may determine whether the patient has the more severe CHARGE phenotype rather than the milder IHH/KS phenotype.

Interestingly, the mutations have been localized to regions around four exons - 2, 6, 8, and 38 - suggesting the possibility of hotspots for IHH/KS mutation. Because of these findings and the absence of nonsense mutations, which often occur in CHARGE syndrome (Vissers, van Ravenswaaij et al. 2004), we provide the first convincing evidence that IHH/KS represents a milder allelic variant of CHARGE syndrome. Although its precise function is uncertain, *CHD7* appears to be important in GnRH and olfactory neuron migration to their embryologic destination in the hypothalamus. *CHD7* is the first chromatin-remodeling protein involved in normal puberty in humans and is the second gene (after FGFR1) identified that results in both normosmic IHH and KS.

3.4. Investigation of mutations in *WDR11*

In *WDR11* the experimental group found six patients with a total of five different heterozygous *WDR11* missense mutations, including three alterations (A435T, R448Q, and H690Q) in WD domains. WD repeats consist on a conserved protein sequence of ≈40 amino acids that typically ends in tryptophan-aspartate. The WD40 protein family comprises many regulatory or adaptor proteins involved in signal transduction, and its members may have one or several WD40 repeats. WD40 mediates protein-protein interaction, such as the β subunits of trimeric G-proteins (Fong, Hurley et al. 1986). WD-containing proteins have 4 to 16 repeating units, which form a circular structure the β propeller (Li and Roberts 2001).

In addition, interactions of *WDR11* with EMX1 were discovered; EMX1 is a homeodomain transcription factor involved in the development of olfactory neurons, and missense alterations reduce or abolish this interaction. The experimental findings suggest that impaired pubertal development in these patients results from a deficiency of productive *WDR11* protein interaction.

3.4.1. Protein Modeling

Since no structure for *WDR11* is available, a model of *WDR11* was constructed based on homology to an A C. elegans Homologue Of Yeast Actin Interacting Protein 1 (AIP1) (PDB code 1NR0) (Mohri, Vorobiev et al. 2004), using the multiple sequence alignment ClustalW. The model was constructed on the region from 70 to 739 and used to investigate the four observed mutations (R395W, A435T, R448Q, H690Q) with respect to their tendency to be detrimental to protein function.

The close structural similarity of the model of *WDR11*, which features two β propeller structures in each protein chain (Voegtli, Madrona et al. 2003) (Fig. 3.3A), with the AIP1 dimer (PDB code 1PGU), indicates that *WDR11* will also be an actin-binding protein (Franke 2004) (Fig. 3.4). *WDR11* contains twelve WD domains, nine (second to tenth repeats) that are confirmed on the basis of direct comparison with the template structure of AIP1 (Fig. 3.3A and 3.3B) and three additional repeats (first, 11th, and 12th) detected by sequence comparison outside the region of the structural model. Like AIP1, *WDR11* is predicted to exhibit two β propellers; WD domains 2–6 are predicted to constitute the first, and WD domains 7–10 are predicted to constitute the second (Fig. 3.3A). The structural model for *WDR11* overlaps well with the known AIP1 structure (Fig. 3.4). This model predicts that *WDR11* has twelve WD domains and that nine of them (second through tenth) participate in the genesis of two consecutive β propellers.

It is known that proteins with repeated WD domains, each of which consists of four antiparallel β strands, form β propeller structures to support interactions with protein-binding partners and to organize and stabilize multiprotein complexes (Higa and Zhang 2007). A β propeller is characterized by 4–8 bladeshaped β sheets arranged around a central axis; each sheet of four antiparallel β strands is twisted so that the first and fourth sheets are close to perpendicular. The last β strand of one WD repeat and the first three β strands of the WD repeat form a blade of the β propeller (Fig. 3.3A).

3.4.2. Analysis of Mutations

Three of the *WDR11* missense mutations leading to R395W, H690Q, and F1150L alter amino acid residues that are completely conserved in all 13 available mammalian and avian orthologs (human, chimpanzee, cow, horse, panda, pig, dog, rat, mouse, rabbit, opossum, chicken, and finch), and a fourth change, A435T, is shared in 11 out of 13 species, suggesting that these substitutions in six independent sporadic patients are very likely to be detrimental. Three of the missense alterations are located directly in the predicted propeller regions of *WDR11*: A435T and R448Q are in the sixth WD domain, and H690Q is in the ninth WD domain (Fig. 3.3B).

In addition we have used SPPIDER to identify possible binding sites in the sequence, as such sites are most likely to affect protein function and signaling. For R395W both POLYPHEN and SIFT predict a possible damaging effect of the mutation. SPPIDER doesn´t identifies this location as a protein interaction site. For A435T all programs give neutral predictions. For R448Q all programs give neutral predictions. For H690Q only POLYPHEN predicts a possible detrimental effect, while both other programs give neutral predictions.

These three mutations are predicted by SPPIDER to alter protein-protein binding domains defined by protein modeling and therefore are likely to disrupt normal protein function (Fig. 3.5). The figure 3.5 shows two prominent external interaction regions (labeled inner and outer surface), where *WDR11* may interact with multiple other proteins, such as EMX1 or actin, as well as an intramolecular interaction region at the interface of the two propeller units.

3.4.3. Discussion

The similarity in structure between *WDR11* and the known structure of AIP1 suggests that *WDR11*, like AIP1, may form a dimer stabilized by interaction with two zinc ions (Fig. 3.6).

The two units, each of which has two β propeller elements, are stabilized by two zinc ions (shown in orange) (Voegtli, Madrona et al. 2003). Zinc ions in proteins are coordinated by a highly conserved set of amino acids, most commonly His, Glu, Asp and Cys (Auld 2001). The experimental structure of 1PGU shows the interaction between zinc and amino acids Asp, His and Glu. Both Zn ions are stabilized by residues coming from different molecules in the dimer and they therefore act as tethers that hold the dimer together.

Because the two protein structures are not identical, deviations arising from their alignment make the position of the Zn ion uncertain to a few angstroms (Fig. 3.4). However, *WDR11* has the required residues (Asp377, Glu384, His501, His508, and Glu510) for zinc binding in the vicinity of the putative zinc position (Fig. 3.7). The R448Q mutation is less than 5 Å from the predicted zinc binding site. Arg residues near Zn coordination sites do not directly interact with the Zn, but they stabilize their environment because Arg is highly positively charged. Replacing Arg with the much smaller Gln residue could influence the zinc-binding propensity of *WDR11* and affect its dimer formation and interactions, including a potential actin interaction predicted by analogy with AIP1 (Kudryashov, Sawaya et al. 2005).

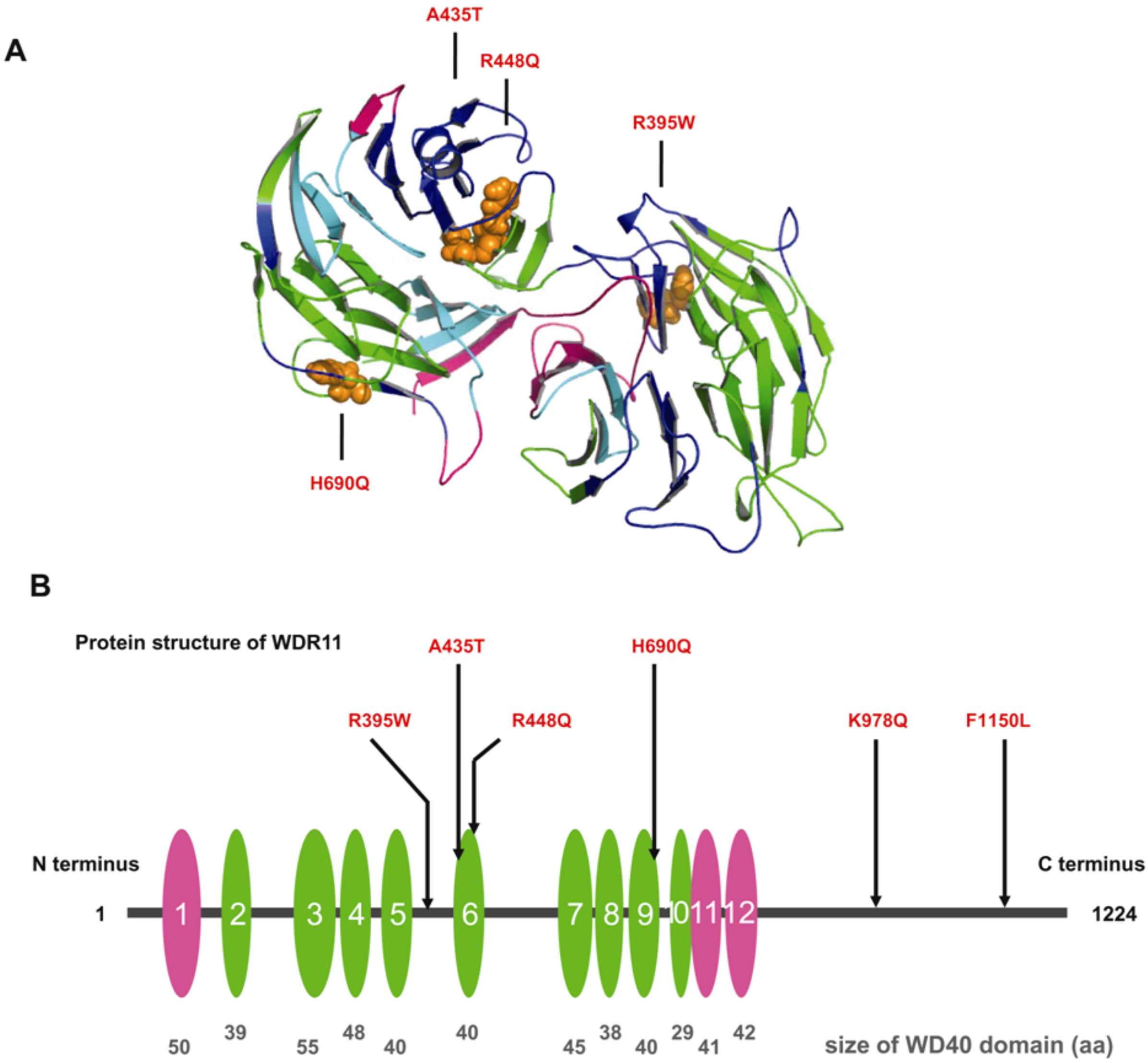


Figure 3.3: *WDR11* Structural Model Indicating the Mutation Sites (A) Model spanning amino acids 70-739 of *WDR11*. *WDR11* forms a double propeller structure, in which the WD domains indicated in (B) form the main structural constituent. Colors correspond to picture B and the sites of the mutations are indicated in orange. (B) Positions of five missense mutations in *WDR11*; WD domains are depicted as ovals. The WD domains predicted on the basis of the model and by SMART are depicted in green and pink, respectively.

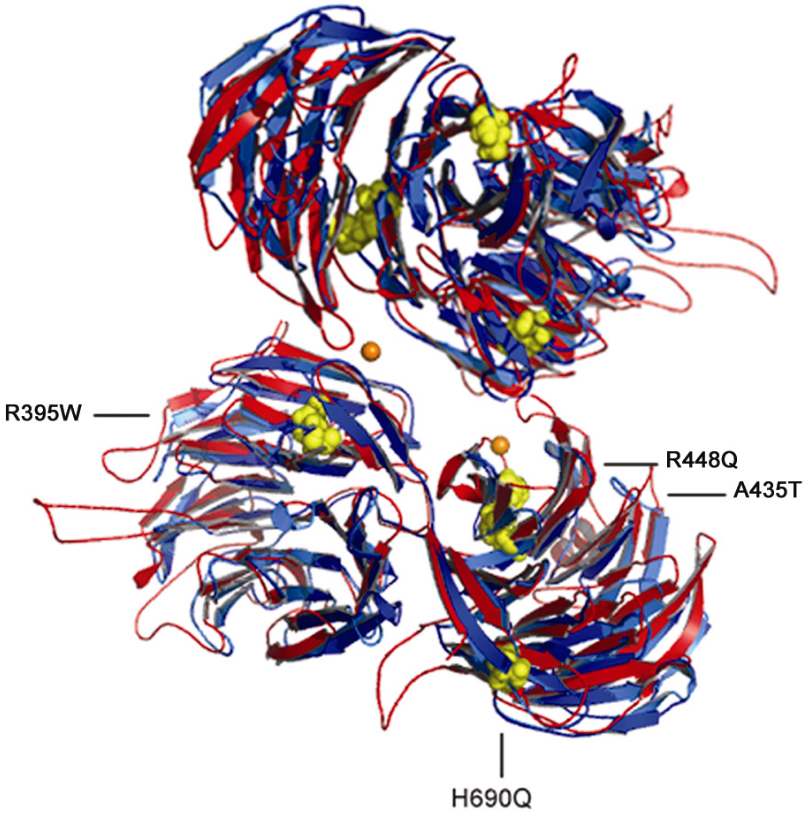


Figure 3.4: Overlay of 1PGU and *WDR11*.
***WDR11* is colored in red and 1PGU in blue, zinc ions correspond to the orange spheres. The sites of mutations in *WDR11* are indicated with yellow spheres.**

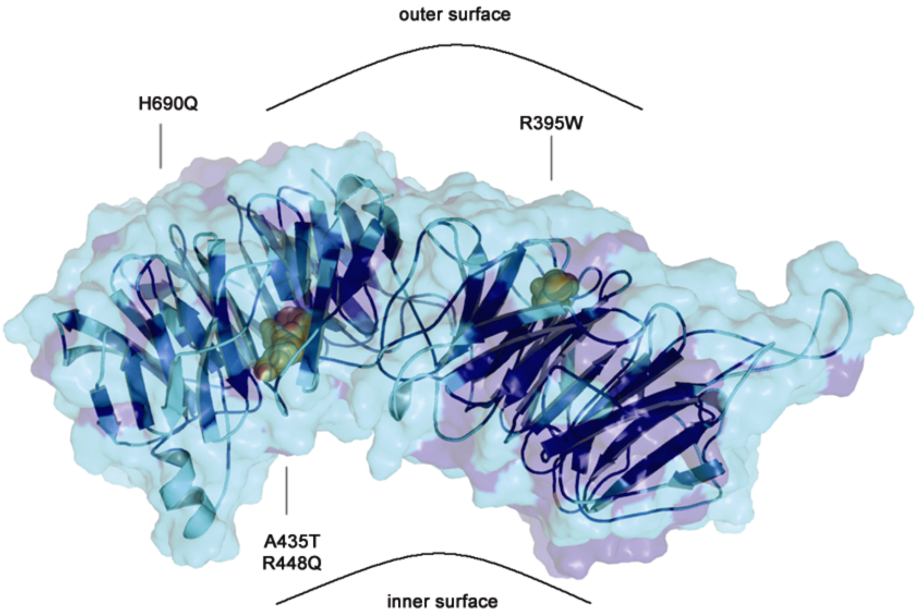


Figure 3.5: Protein-protein interaction region of *WDR11*.
We used SPPIDER to detect protein-protein interaction regions are colored in bright cyan, in the model for *WDR11* non-interacting regions are in dark blue, here shown in a view obtained by rotating panel A 90 degrees out of the plane on a horizontal axis.

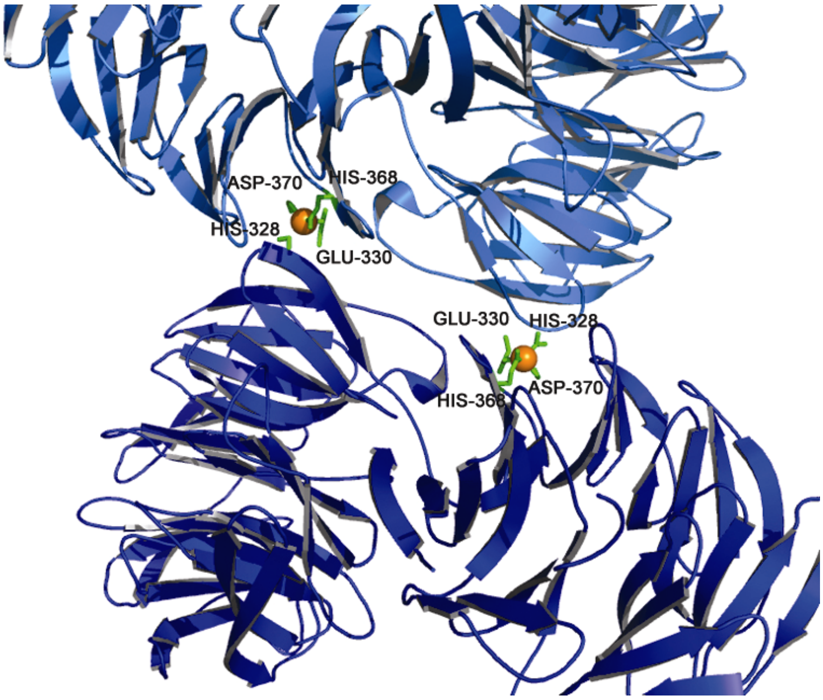


Figure 3.6: Two zinc ion binding sites of 1PGU.

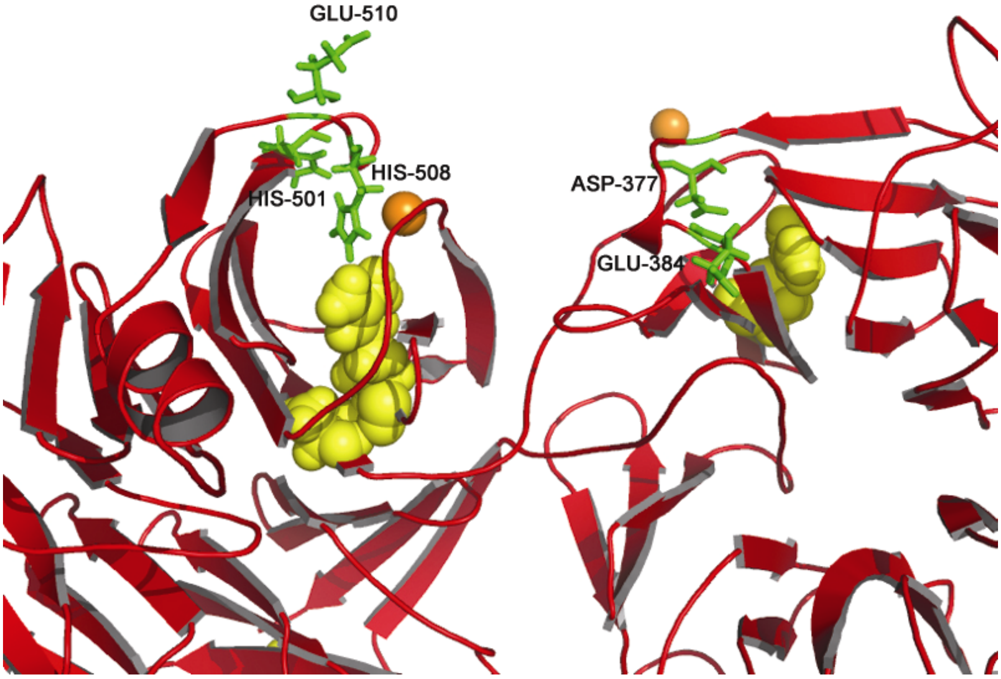


Figure 3.7: Glu, His and Asp residues in the model of *WDR11* in the vicinity of the proposed zinc position based upon 1PGU.
A single unit of the proposed dimer is shown.

Taken together, the genetic and functional data work performed for *WDR11* provide strong evidence for missense sequence variants of *WDR11* as a cause of IHH and KS in a proportion of cases of this genetically heterogeneous condition. This adds to the growing list of genes known to be mutated in IHH and KS and will open new investigative routes for understanding the development of normal human puberty and reproduction.

3.5. Investigation of mutations in Nasal Embryonic LHRH Factor (*NELF*)

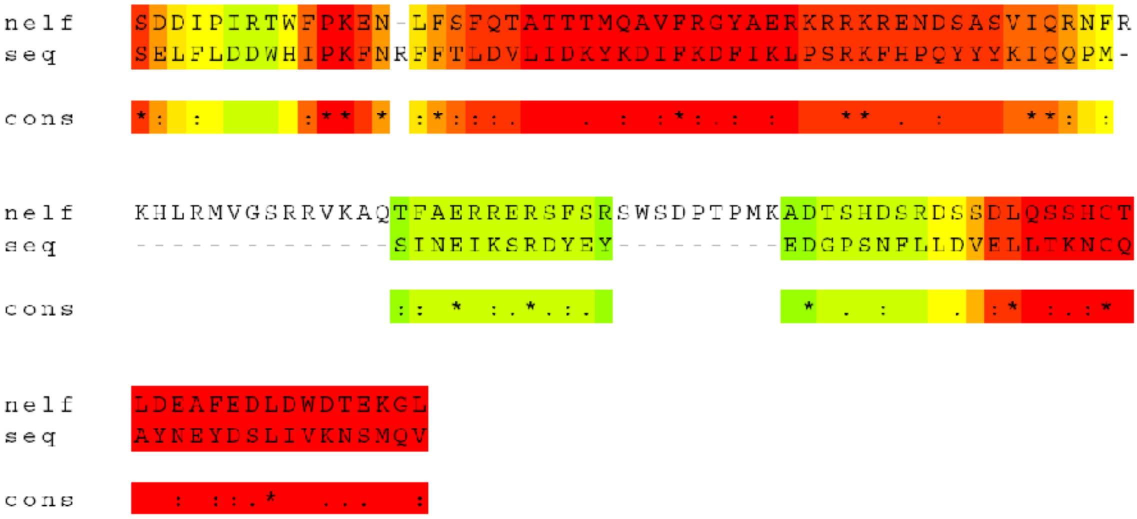
Recently, nasal embryonic LHRH factor (*NELF*) was identified from a differential screen in migratory GnRH neurons, suggesting a potential role in KS (Kramer and Wray 2000; Kramer and Wray 2001). Only two *NELF* mutations have been previously reported in humans with nHH/KS. One heterozygous *NELF* mutation was identified in a nHH patient, but neither *in vitro* analysis or mutation screening of additional nHH/KS genes was performed (Miura, Acierno et al. 2004). Another KS male had both heterozygous *NELF* and *FGFR1* mutations, suggesting a digenic pattern (Pitteloud, Quinton et al. 2007). Additionally, multiplex ligation dependent amplification (MLPA) analysis in 100 nHH/KS patients did not reveal any heterozygous *NELF* deletions (Pedersen-White, Chorich et al. 2008). To date, no *NELF* mutations, backed by functional analysis, have been reported in monogenic nHH or KS. Therefore, the purpose of the present study was to determine whether *NELF* mutations occur in monogenic nHH/KS or if they are involved in a new digenic pattern.

In this investigation one heterozygous missense mutation in *NELF* at position 253 changing from Ala to Thr was investigated. This mutation affects a completely conserved Ala253 residue in frog, chicken, mouse, rat, dog, and human illustrating its structural and functional importance. Three of 168 well-characterized nHH/KS patients (Bhagavath, Podolsky et al. 2006) demonstrated *NELF* mutations. One is the first identified human *NELF* mutation, supported by in vitro analysis (protein reduction), in a nHH patient without mutations in 11 other known nHH/KS genes. In addition, two different mutations are involved in new digenic patterns of *NELF*/KAL1 and *NELF*/TACR3 in two unrelated KS patients. The *NELF* mutations showed in vitro evidence as reduced protein expression, splicing defects, or altered nuclear localization. Findings from the experimental group suggest that *NELF* is associated with both nIHH, in which no additional known nIHH/KS gene mutations are identified, and KS when mutations in additional genes are present. No *NELF* mutations have been reported in monogenic nHH or KS. Therefore, the purpose of the present study was to determine whether *NELF* mutations occur in monogenic nHH/KS or if they are involved in a new digenic pattern.

3.5.1. Protein Modeling

A model for the N-terminal region (209-508) of the protein was constructed based on homology to an acetylated Rsc4 tandem bromodomain Histone Chimera (PDB code 2R10) (VanDemark, Kasten et al. 2007), using multiple sequence alignment ESPRESSO (Armougom, Moretti et al. 2006). Another model was constructed for the extended N-terminal region of *NELF* model from amino acid 209 to 528 (including the mutation A253T) shown in figure 3.9B.

I then used MOE to generate a model on the basis of the alignment, which has low confidence for the C-terminal region of the sequence modeled, but high confidence (Fig. 3.8) in the vicinity of the region.



```
nelf  SDDIPIRTWFPKEN-LFSFQTATTTMQAVFRGYAERKRRKRENDSASVIQRNFR
seq   SELFLDDWHIPKFNRFFTLDVLIDKYKDIFKDFIKLPSRKFHPQYYYKIQQPM-

cons  *: :     :** * :*:::.   . : :*:.: :   ** . :    **: :

nelf  KHLRMVGSRRVKAQTFAERRERSFSRSWSDPTPMKADTSHDSRDSSDLQSSHCT
seq   --------------SINEIKSRDYEY---------EDGPSNFLLDVELLTKNCQ

cons                :: * :.*.:.           * . :   . :* :.:*

nelf  LDEAFEDLDWDTEKGL
seq   AYNEYDSLIVKNSMQV

cons    : ::.*  ...  :
```

Figure 3.8: Quality assessment of the degree of homology in the 70 amino acid N-terminal region of *NELF*. Red regions indicate regions of high homology, while yellow and green regions indicate low homology. Dashes indicate insertions in the *NELF* sequence not present in the template.

While the homology in the 70 amino acid N-terminal region demonstrated in figure 3.9A permits the construction of a detailed model, the resolution of the C-terminal region of the fragment is likely to be lower as a consequence of lower homology. However, the overall model is sufficient to demonstrate that the C-terminal region of the structure has no intramolecular contacts with the N-terminal region where the mutation Ala253Thr (highlighted in magenta) occurs, such that the effects and exposure of the mutation can be investigated on the basis of the model for the 70 amino acid model shown in figure 3.9A.

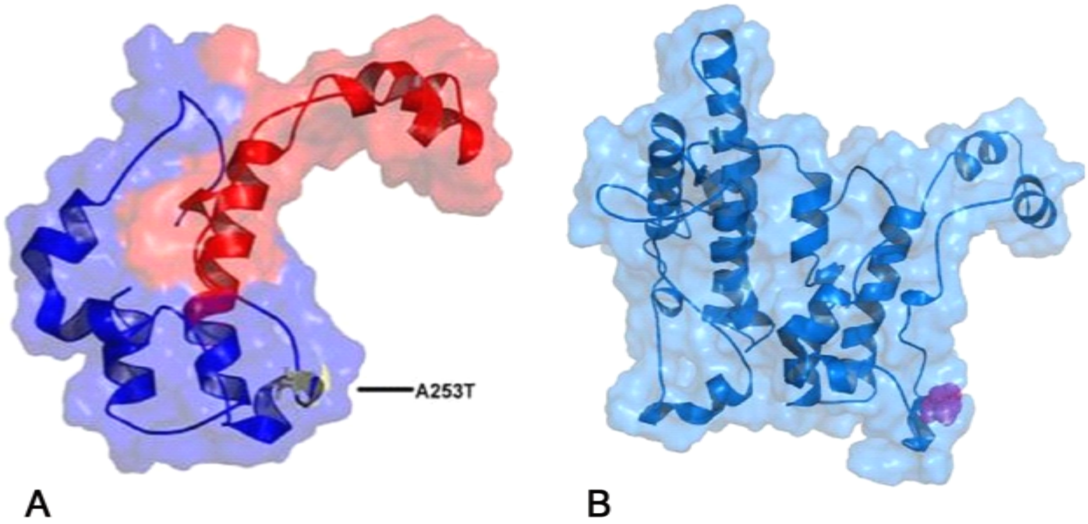


Figure 3.9: (A) Model of N-terminal region of *NELF* spanning amino acids 209-331 obtained by alignment to 2R10. Blue regions in the model correspond to high-confidence regions based on the quality alignment to the template, while red regions correspond to low-confidence regions. The site of the mutation, which is located at the outside of the protein withing a high-confidence region, is indicated in yellow. (B) Model for the extended N-terminal region of *NELF* model from amino acid 209 to 528. A253T mutation is highlighted in magenta.

3.5.2. Conclusions

We have investigated the observed mutation A273T with respect to its tendency to be detrimental to protein function using bioinformatics based methods (POLYPHEN and SIFT). In addition we have used SPPIDER to identify possible binding sites in the sequence, as such sites are most likely to affect protein function and signaling. SIFT predict a possible damaging effect of the mutation, while POLYPHEN gives a neutral prediction. In addition SPPIDER identify this location as a protein interaction site (Fig. 3.10).

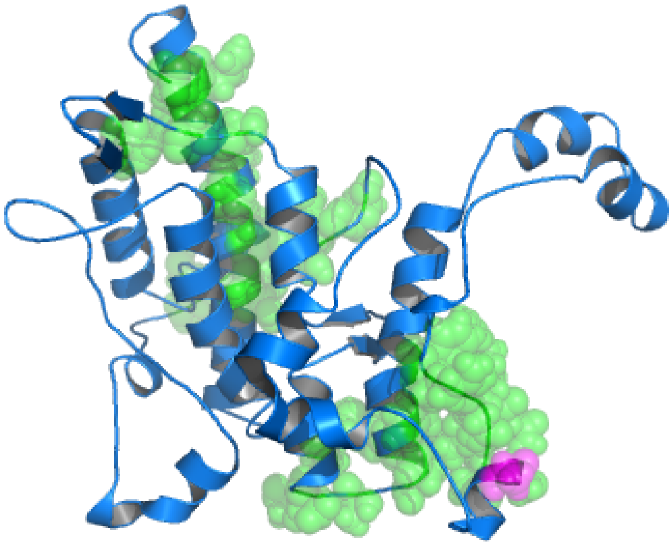
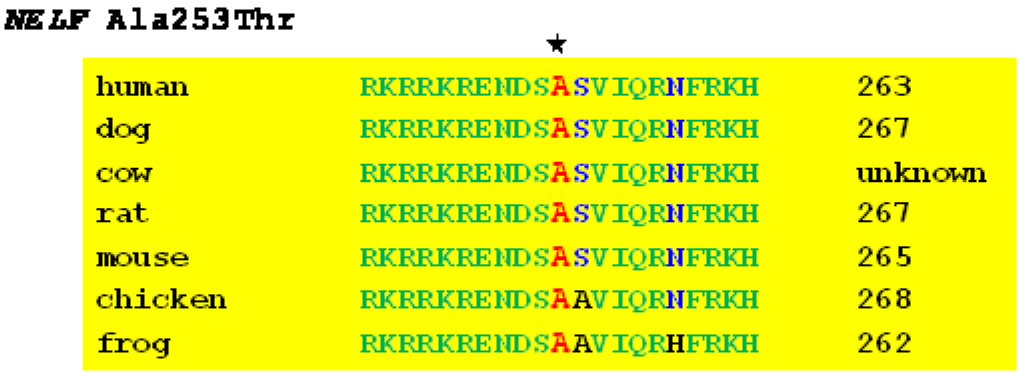


Figure 3.10: The observed mutation (shown in magenta) occurs in the central region of an extended protein-protein interaction region highlighted in green on the model of the 209-528 amino acid N-terminal of *NELF*.

Experimental and theoretical data indicate that *NELF* is highly likely to be causative in nHH and KS. The heterozygous c.757G>A (p.Ala253Thr) *NELF* missense mutation in the patient resulted in a drastic amino acid substitution from a hydrophobic to a hydrophilic residue. Since this variant was not identified in 372 controls and protein expression was markedly reduced in vitro, this is highly likely to represent a mutation. Further support comes from complete conservation of the Ala253 residue in all seven species identified (Fig. 3.11), as well as findings from both SIFT (sorting intolerant from tolerant) and protein modeling.



NELF Ala253Thr

human	RKRRKRENDSASVIQRNFRKH	263
dog	RKRRKRENDSASVIQRNFRKH	267
cow	RKRRKRENDSASVIQRNFRKH	unknown
rat	RKRRKRENDSASVIQRNFRKH	267
mouse	RKRRKRENDSASVIQRNFRKH	265
chicken	RKRRKRENDSAAVIQRNFRKH	268
frog	RKRRKRENDSAAVIQRHFRKH	262

Figure 3.11: Ala253 in *NELF* is evolutionarily fully conserved in seven available orthologs. Thus the change of this hydrophobic residue to a hydrophilic Thr is likely to be deleterious. This has also been shown by SIFT.

3.5.3. Discussion

In summary, a large number of nHH/KS patients were screened to determine the prevalence of NELF mutations. Three novel human *NELF* mutations were identified in three unrelated nHH/KS patients, as well as three unreported *NELF* rare sequence variants of undetermined significance. The prevalence of 1.8% (3/168) suggests that mutations of *NELF* are involved in the pathogenesis of a subset of nHH and KS cases. We believe that the compound heterozygous sequence variants are the first reported *NELF* mutations, confirmed by reduced protein expression in vitro, in a nHH patient who does not possess mutations in any of the other common nHH/KS genes.

In addition, the previously unreported digenic association of *NELF*/KAL1 and *NELF*/TACR3 was identified in two unrelated KS patients. These findings suggest that mutation of one *NELF* allele may not be sufficient to result in disease unless there is an additional *NELF* allele or coexistent mutation in another gene, such as KAL1 or TACR3. These results, along with two previous reports of *NELF*/FGFR1 (Pitteloud, Quinton et al. 2007) and PROKR2/KAL1 (Dode, Teixeira et al. 2006), highlight the importance of biallelic mutations in nHH and KS and warrant consideration of mutation screening of known genes to add to the growing list of digenic patterns in nHH and KS.

The human *NELF* gene has been localized to chromosome 9q34.3. A 9q34.3 deletion syndrome has been described, in which two males had abnormal genitalia and a female displayed anosmia (Yatsenko, Cheung et al. 2005). The deletion encompasses *NELF*, thereby implicating *NELF* in the partial KS phenotypes of these individuals. Two *NELF* mutations identified in this study showed reduced protein expression and one demonstrated a splicing defect leading to protein truncation. Based upon our findings, as well as the whole *NELF* deletion contained within the 9q34.3 contiguous gene deletion syndrome, we propose that haploinsufficiency of *NELF* is associated with both monogenic nHH and digenic KS.

3.6. Summary

Advances in the sequencing of the DNA resulted in an explosion of sequence information (Linnarsson 2010), but much of this information cannot be directly related to biological function. One important avenue to elucidate the biological function of a protein, or the relevance of specific mutations, is the study of its three dimensional structure. Today, however, there is still an enormous gap between the number of sequences and the number of protein structures. There are still thousands of biologically important proteins waiting to be crystallized an effort much more involved that the state-of-the-art for sequencing. For these reason computational methods, such as homology modeling, docking, molecular dynamics, to name a few, are important to help experimentalist in their work.

Advances in computational power and innovation have led to the development of novel and accurate methods for the 3-D modeling of proteins (DiMaio, Terwilliger et al. 2011). Some of the new methods have been proven to be accurate and rapid. However, the capability to be able to build a protein model very close to the native structure of the protein reliably is a challenging assignment and is still under development (Floudas, Fung et al. 2006). When homology to known proteins is high, accurate predictions of protein modeling assist in understanding the mechanism of action of proteins and even aid in drug design. Considering all the developments in the field, the task seems to be achievable in the near future.

For proteins with little homology to structurally resolved protein, the situation is more complex. Presently available methods only result in low resolution models which can be used only indirectly for others applications. Nevertheless even such models help experimentalist on the basis of a low resolution model to formulate hypothesis regarding protein function that may be experimentally verified. Protein structure prediction can guide mutagenesis experiments, or make hypothesis on relations about structure-function; the conserved regions of the protein could be used to identify putative active sites and binding pockets.

In this chapter I have reported work of the identification causes for Kallman syndrome/idiopathic hypogonadotropic hypogonadis, where the identification of mutations in candidate genes assumes particular importance toward understanding the molecular basis of both nHH and KS. For *CHD7* we hypothesized that the gene would be involved in the pathogenesis of IHH and KS (IHH/KS) without the CHARGE phenotype and that IHH/KS represents a milder allelic variant of CHARGE syndrome. *CHD7* represents the first identified chromatin-remodeling protein with a role in human puberty and the second gene to cause both normosmic IHH and KS in humans. We were able to generate a model for the 3-D structure, for which no experimental structure was available, to observe possible functional effects of the mutations in the patients exhibiting the syndrome. Our findings indicate that both normosmic IHH and KS are mild allelic variants of CHARGE syndrome and are caused by *CHD7* mutations. For *WDR11* six patients where found with a total of five different heterozygous *WDR11* missense mutations, including three alterations (A435T, R448Q, and H690Q) in WD domains important for β propeller formation and protein-protein interaction. I have also generate a model for *WDR11* where our findings suggest that impaired pubertal development in these patients results from a deficiency of productive *WDR11* protein interaction. Finally for *NELF* we wanted to determine if mutations in *NELF*, a gene isolated from migratory GnRH neurons, cause normosmic idiopathic hypogonadotropic hypogonadism (IHH) and Kallmann syndrome (KS). Studies on patients with *NELF* mutations where confirmed. *In vitro* evidence of these *NELF* mutations included reduced protein expression and splicing defects. Our findings suggest that *NELF* is associated with normosmic IHH and KS, either singly or in combination with a mutation in another gene.

The results obtained in this chapter demonstrate that methods of protein structure prediction are increasingly able to contribute to and complement experimental efforts to elucidate the function of genes for which structural information for the corresponding proteins is unavailable. Due to the rise of the number of structures in the PDB data base there is an increasing likelihood that at least some regions of the protein in question have high homology with a known protein which helps in classifying the function and generates a structural template that may be used by other, bioinformatics-based methods to elucidate the role of the protein and the relevance of mutations observed in specific patients. In addition prediction of protein structure is of growing importance even in absence of high homology, as it can offer some insights into function. The models generated by present-day methods in the case of low homology are overall of low quality, such that an analysis based on such models may be used to motivate additional experiments or motivate further modeling efforts, but cannot generate sound conclusions on their own. In this context the development of methods for absolute quality assessment, which would tell us to which degree we can rely on low-resolution models, would be a significant step to increase the usefulness of methods for protein structure prediction.

4. Protein Protein Interactions

4.1. Introduction

Biomolecular interactions have an important role in controlling biological processes at the protein-protein, protein-DNA, DNA-small molecule and protein-small molecule level. Their identification, quantification, and control are key to areas such as drug-discovery, cell signaling, elucidation of biosynthetic pathways and understanding of enzyme catalysis. Strategies to modulate protein-protein binding by rationally designed, competitive small molecules (including peptides) have emerged as a promising avenue for drug discovery (Wells and McClendon 2007).

Recent studies discovered small molecules that bind to 'hotspots' on the contact surfaces involved in protein-protein interactions. For example: *Bcl-XL binders* are members of the B-cell lymphoma 2 (Bcl-2) families and are important regulators of apoptotic cell death (Sattler, Liang et al. 1997; Adams and Cory 1998). The binding of Bcl-2 and Bcl-XL with BAK and BAD (important in the treatment of cancer) inhibits apoptosis. Recently, compounds that bind to the hydrophobic helical domain of Bcl-XL, Bcl-2 and another anti-apoptotic molecule, Bcl-W were synthetized, where the most potent compound was found to be ABT-737 (Oltersdorf, Elmore et al. 2005). These small molecules do not bind well to other anti-apoptotic members of the Bcl-2 family such as MCL1 (myeloid cell leukaemia sequence 1) and Bcl-B46. ABT-737 binds to the same region of Bcl-XL, but not closely mimics the atomic details of the peptide. Instead, the small molecule traps a slightly different conformation of Bcl-XL, binding in deeper cavities with more folded grooves.

Another example are *TNF disruptors* (cytokine tumour-necrosis factor), which are key factors in inflammatory responses. Because there is also an important role of TNF in treating arthritis it is not surprising that there is considerable interest in developing small molecules or peptides that can disrupt the interaction between TNF and its receptors, TNFR1 and TNFR2. A class of small molecule that targets TNF was discovered, by using fragment screening (He, Smith et al. 2005). These molecules disrupt TNF by binding and displacing one of the three monomers that constitute TNF. Because of their moderate affinities, small molecules of this class are not seriously considered as drug candidates; however, their discovery shows that even interfaces in oligomeric proteins can bind to small molecules. Many progresses have been made in the past years, however for the regulary discovery of inhibitors many more promising targets exist.

As a final example *IL-2 binders* (cytokine interleukin-2) have a key role in the activation of T cells and in the rejection of tissue grafts. A series of small molecules that bind to IL-2 were synthetically produced and by testing them showed that the binding surface on IL-2 is adaptive and can bind to a small molecule with high affinity using the same main hotspot residues. In collaboration with the experimental group of Dr. Katja Schmitz at the KIT (Karlsruhe Institute of Technology) I have investigated chemokine interactions with their receptors in order to develop peptides that may interfere with this interaction mechanism. Chemokines, or chemotactic cytokines, are small signalling proteins that guide the migration of cells along a concentration gradient. As such, they are involved in a number of important pharmaceutically relevant processes, including inflammation and angiogenesis (Graves and Jiang 1995). Chemokines perform their biological function through interaction with chemokine receptors, which belong to the family of intensely studied G-protein coupled receptors (GPCRs). GPCR's, which have been notoriously difficult to characterize structurally (rhodopsin and human B2-adrenergic GPCR (Rasmussen, Choi et al. 2007) are two of the few examples of characterized structures), have been one of the main targets of drug discovery over the last 15-20 years: over 30% of the drugs available on the market today interfere with GPCR function (Hill 2006).

In this chapter I have used a biophysical simulation approach, described in section 4.2, to investigate the binding hot-spots and affinity of peptides mimicking the unknown receptor for several structurally resolved chemokines. Specifically we have first used an all atom free-energy approach to model the binding of the chemokine CXCL8 (inteleukin-8) to the N-terminal peptide of its cognate receptor CXCR1 based on existing NMR data. Individual amino acid residues have been substituted by alanine and the resulting change in binding energy was analyzed to find out which residues of the N-terminus of the receptor make the largest contribution to the binding energy. These results increase our understanding of the dominant contributions to the free energy of protein-protein interactions and can guide experiments aimed at the design of protein interaction inhibitors.

We then performed a computational alanine screen investigating mutations for a fragment of membrane protein ERBIN in complex with the cytosolic domain of ERBB2, a receptor tyrosine kinase present on the basolateral membrane of polarized epithelia with important functions in organ development and tumorigenesis (Section 4.4). Finally I extend the work on CXCL8 and its receptor CXCR1 to the CCR3 receptor to predict the effect of all possible mutations on the previously identified hotspots of the extracellular domain that are crucial for eotaxin binding (Section 4.5). We found out by these studies which residues of the receptor make the largest contribution to the binding energy.

Based the atomistic analysis of the protein-protein interaction profiles and the findings from binding experiments we will design novel peptides that bind competitively to the chemokines. Elucidating the binding motifs and locating their position on a three-dimensional model of the receptor reduces the experimental effort to investigate chemokine receptor recognition and activation.

4.2. Methods

In the investigations reported in this chapter we have used a biophysical simulation methodology that combines Monte-Carlo simulations for the conformational space with configuration based estimates of the binding affinity, similar to the methods discussed in section 2.5. The POEM (protein optimization with energy methods) (Anfinsen 1973; Herges and Wenzel 2004; Verma, Gopal et al. 2007; Verma and Wenzel 2009) method is based on Anfinsen´s hypothesis, which states that most proteins are in thermodynamic equilibrium with their environment in their native state (Anfinsen 1973). For most proteins the native conformation corresponds to the global optimum of the free energy of the protein. This global optimum of a complex energy landscape can be obtained with high efficiency using stochastic optimization methods mapping the folding process found in nature onto a non-equilibrium dynamical process that explores the free-energy surface of the protein. Using this method we locate not only the native conformation (Schug and Wenzel 2004; Gopal and Wenzel 2006; Schug and Wenzel 2006; Verma, Gopal et al. 2007; Verma, Gopal et al. 2008) but typically characterize the entire low-energy ensemble of competing metastable states. Based on the underlying atomistic biophysical forcefield, these methods can be used to decompose the interaction energy of protein complexes into the contribution of individual residues and account for structural relaxation of the binding partners after mutation (Herges and Wenzel 2004; Verma, Patil et al. 2008).

4.2.1. The forcefield PFF02

PFF02 is an all-atom (with exception of apolar CHN groups) free-energy forcefield, which was initially developed to describe protein folding and large scale conformational change (Verma and Wenzel 2009) and thus is intended for simulations in which bond-length and chemically defined bond angles do not change. Similar to other biomolecular forcefields PFF02 implements the following non-bonded interactions. Since hydrogen bonding and solvent interaction are the two major contributions to protein folding, these interactions are specially emphasized and modeled by two contributions specifically adapted to protein data in PFF01/02.

- Lennard-Jones: the van der Waals interactions and a repulsive terms are included in the forcefield as a Lennard-Jones 6-12 potential.

$$V_{lj}(\vec{r}) = V_0 \sum_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - \left(\frac{2R_{ij}}{r_{ij}} \right)^{6} \right]$$

where i, j represent the atoms included in the forcefield, r_{ij} is the distance between these atoms and R_{ij} are the Lennard-Jones radii ($R_{ij} = R_{ii}\ R_{jj}$). The parameters for the Lennard-Jones potential were derived as a potential of mean force from experimental data by fitting short-range (2 Å-5 Å) radial distributions of a set of 138 different proteins which are believed to span a wide range of different folds (Avbelj and Moult 1995).

- Electrostatics: The electrostatic interaction is modeled using a columbic potential augmented by group specific dielectric constants accounting for local polarizibility. This term is split into main chain and side chain contributions.

$$V_{ele}(\vec{r}) = V_{main}(\vec{r}) + V_{side}(\vec{r}) = \sum_{ij} \frac{q_i q_j}{\varepsilon_{g(i)g(j)} r_{ij}}$$

where i, j represent the atoms included in the forcefield, q_i and q_j are the corresponding partial charges, r_{ij} is the distance between these atoms and $\varepsilon_{g(i)g(j)}$ are group-specific dielectric constants.

- Hydrogen bonding: Generally hydrogen bonding is not explicitly modeled, but its contributions are embedded partly in the electrostatics and Lennard-Jones. PFF02 differs from other biomolecular forcefields in recognizing that the angle dependence of hydrogen bonds is not well described using only such terms an additional term to describe hydrogen bonding was added. The electrostatic interaction between the dipoles of the amino and carboxylgroups of the mainchain is given by

$$V_{hbdipole} = \frac{0.1064e^2}{4\pi\varepsilon\varepsilon_0} \left(\frac{1}{r_{C_iH_j}} - \frac{1}{r_{C_iN_j}} - \frac{1}{r_{O_iH_j}} + \frac{1}{r_{O_iN_j}} \right)$$

(where i, j counts the amino acids with i belonging to the carboxyl- and j the amino group, e equals one elementary charge, r_{XiYj} gives the distance of the atoms X from amino acid i and Y

from amino acid *j*). An additional short-ranged term which corrects the hydrogen bonding by considering the alignment of the hydrogen bond with respect to the donor and acceptor groups (Sipple 1984).

$$V_{hbcorr} = V_0 \sum_{ij} R(r_{H_iO_j})\Lambda(\alpha_{ij}, \beta_{ij})$$

where $V_0 = -2.12$ kcal/(mol A), α is the NHO angle, β the angle between the CO and NH-dipoles, *R(r)* gives the radial and *Λ(α)* the angular dependence to the correction potential. *R(r)* and *Λ(α, β)* are defined as

$$\begin{aligned} R(r) &= s_{2.4,0.075}(r) \\ \Lambda(\alpha,\beta) &= s_{45,5}(\alpha)s_{40,5}(\beta)s_{1.5,0.05}\left(\sqrt{\frac{\alpha}{30}^2+\frac{\beta}{24}^2}\right)^2 \text{ where} \\ s_{A,B}(x) &= \frac{1}{2}\left(1-tanh\left(\frac{x-A}{B}\right)\right) \end{aligned}$$

The hydrogen bonding term in PFF02 interpolates by these contributions.

$$V_{hb} = \lambda V_{hbdipole} + (1-\lambda)V_{hbcorr}$$

where λ gives the strength of correction between [0, 1] with $\lambda = 1$ meaning that the hydrogen bonding is modeled by pure dipole-dipole interaction. In PFF01/02 the value of λ is 0.75.

• Solvation: The solvent energy and entropy influences the conformational changes and binding of proteins and contributes to the free-energy of the system. On the surface of a protein there are important solvent effects: the hydrophobic effect, i.e. changes in entropy of water molecules depending on the exposition to apolar groups, changes in the conformational entropy of the protein sidechains and the modulation of the solvation of charged side groups. The solvation effects are modeled in PFF02 via an implicit solvent model based on the Solvent Accessible Surface Area (SASA) of the protein (Lee and Richards 1971). The SASA is calculated by rolling a water sphere of radius 1.4 Å over the protein surface which is defined by the Lennard-Jones radii. The solvation term is given by the relation

$$V_{sol} = \sum_i \sigma_{PT(i)} A(i)$$

where *PT(i)* is the potential type of atom *i*, $\sigma_{PT(i)}$ describes the Atomic Solvent Parameter (ASP) according to the potential type and *A(i)* is the SASA of the atom *i*. The parameters are derived by fitting first a SASA model (Eisenberg, Wilcox et al. 1986) to reproduce the enthalpies of solvation of the tripeptides Gly-X-Gly (Sharp, Nicholls et al. 1991) and then adjusting these parameters to stabilize the native structure of the villin headpiece (Herges and Wenzel 2004).

- Local electrostatics correction: Amino acids have a preference for secondary structure elements. For example tryptophan, threonine occurs mostly in β sheet regions of the Ramachandran plot, whereas alanine prefers α-helical region. These preferences are influenced by different electrostatic interactions of the main chain dipoles in their local environment (Avbelj and Moult 1995). The interaction E_{local} is defined as the electrostatic energy of the mainchain CO and NH groups of a residue arising from interactions with the main chain CO and NH groups within that residue and with the adjoining peptide groups. So NH_i interacts with CO_{i-2}, NH_{i-1}, CO_i & NH_{i+1} and CO_{i-1}, NH_i, CO_{i+1} & NH_{i+2} interact with CO_i.

$$V_{\mathrm{local}} = \lambda_{\mathrm{local}} \frac{332.150625 \times \zeta}{2} \sum_j \sum_i \frac{q_i q_j}{r_{ij}}$$

where q_i is the charge on the atom and r_{ij} is distance between the atoms. The parameter ζ is an amino acid specific parameter.

- Torsional energy: A weak dihedral angle dependent energy term is introduced to stabilize the residues in the β sheet regions of the Ramachandran plot. The interaction has a favorable energy contribution of 0.8 kcal/mol (maximum) for the residues forming β sheet.

$$V_{\mathrm{tor}} = \lambda_{\mathrm{tor}} \sum_i e^{\gamma_\phi (\phi_i - \phi_0)^2 + \gamma_\psi (\psi_i - \psi_0)^2}$$

4.2.2. Simulation Protocol

Much of the work in the following chapter focuses on the study of the effect of mutations on the binding of proteins and peptides. Mutations were modeled by removing all side chain atoms beyond the carbon and attaching a hydrogen atom in their place. In docking simulations we generated 5000 random structures by moving the peptide center of mass at least 15 Å away from the docking site. In none of these starting conformations the peptide was in contact with the chemokine. For each of these decoys we then performed Monte-Carlo simulations (Binder 1996). This stochastic approach explores the energy landscape by random changes in the geometry of the molecule. In this way, a large area of the configurational space is searched. By using an appropriate acceptance criterion thermodynamic properties of the system can be evaluated. Monte-Carlo is not a deterministic method and does not describe time evolution of the system in a form suitable for viewing, but is well suited to rapidly converge thermodynamic properties because its individual step is much large than that of molecular dynamics simulations (Leach 2001). The simulations performed here comprised center-of-mass moves with a random displacement of 0.5 Å and a rigid body rotation around the center of mass of the peptide by at most 3°. Variations of the dihedral angles of the backbone and side chains are considered in the simulation, while intramolecular degrees of freedom (i.e. all other angles and bond lengths) are kept constant. For subsequent analysis we then decomposed the interaction energy of the two protein complexes into the contribution of individual residues.

4.3. Elucidating specificity for interleukin-8 and its receptor CXCR1

The first part of this set of investigations centered on the complex of the chemokine interleukin-8 (CXCL8) and the N-terminal peptide of its cognate receptor CXCR1 (Fig. 4.1) were used (PDB code 1ILP). The aim of this study in collaboration with Dr. Katja Schmitz is to find small molecules that bind to chemokines and thus prevent the activation of the receptors. In the long run therapies for chronic inflammations, autoimmune diseases and allergies may be developed based on these inhibitors. For CXCL8/CXCR1 we performed a computational alanine scans based on existing structural data.

Interleukin(IL)-8 is a member of the C-X-C family of chemokines produced by macrophages and other cell types such as epithelial cells. Apart from neutrophils and monocytes (Gerszten, Garcia-Zepeda et al. 1999), numerous sessile cell types have been shown to express IL-8 receptors. These cell types include neurons (Horuk, Martin et al. 1997), various cancer cells (Norgauer, Metzner et al. 1996; Miller, Kurtzman et al. 1998; Metzner, Hofmann et al. 1999), and endothelial cells (Murdoch, Monk et al. 1999). CXCL-8 shows high-affinity binding to the CXCR1 (IL-8 receptor type 1) and the CXCR2 (IL-8 receptor type 2). Although the CXCR1 is selectively activated by IL-8 only, the CXCR2 responds to several additional chemokines including growth-related protein-α (GROα), neutrophil-activating peptide-2, and epithelial-derived neutrophil attractant-78. This chemokine is one of the major mediators of the inflammatory response such as for example gingivitis (Huang, Haake et al. 1998) and psoriasis (Utgaard, Jahnsen et al. 1998).

We used the all-atom free-energy modeling approach described in section 4.2 to calculate the binding energy of the complex and to decompose the interaction energy into the contributions of individual amino acids. To identify amino acids with important contributions to the total binding energy we conducted an *in silico* alanine mutation screen of the receptor-peptide (Table 4.I).

Residue No.	Aminoacid	Solvation	Side chain electrostatics	Main chain electrostatics	Backbone torsion	Lennard-Jones	Total
1	Met	0,90	-2,82	-0,30	0,09	0,48	-1,65
2	Trp	0,85	-1,21	-0,29	0,09	0,50	-0,06
3	Asp	0,90	**6,14**	-0,30	0,09	0,48	**7,31**
4	Phe	**7,79**	-0,05	-0,24	0,09	0,84	**8,43**
5	Asp	1,02	**6,95**	-0,54	0,03	0,59	**8,05**
6	Asp	-1,01	**6,11**	-0,30	0,09	0,62	**5,51**
7	Gly	-0,52	0,29	-0,30	0,09	0,45	0,01
8	Met	3,07	0,29	-0,30	0,09	0,59	3,74
9	Pro	2,09	0,25	-0,31	0,09	0,67	2,79
10	Pro	2,07	0,20	-0,23	0,09	0,78	2,91
11	Asp	1,09	9,42	-0,30	0,09	0,53	**10,83**
12	Glu	1,29	6,23	-0,30	0,09	0,72	**8,03**
13	Asp	0,90	7,14	-0,30	0,09	0,48	**8,31**
14	Tyr	4,15	-0,07	-0,30	0,09	0,59	4,46
15	Ser	1,28	0,29	-0,30	0,09	0,50	1,86
16	Pro	2,99	0,25	-0,26	0,09	0,69	3,76

Table 4.I: Energy contributions (in Kcal/mol) of alanine substitutions of individual amino acids in the N-terminal peptide of CXCR1. The total energy is split into contributions by solvation effects, electrostatics side and main chain, backbone torsion, and Lennard-Jones potential. Large changes in interaction energy can be attributed to either solvation effects in case of polar and nonpolar residues or to electrostatic effects for charged side chains. Bold typeface indicates stronger interactions.

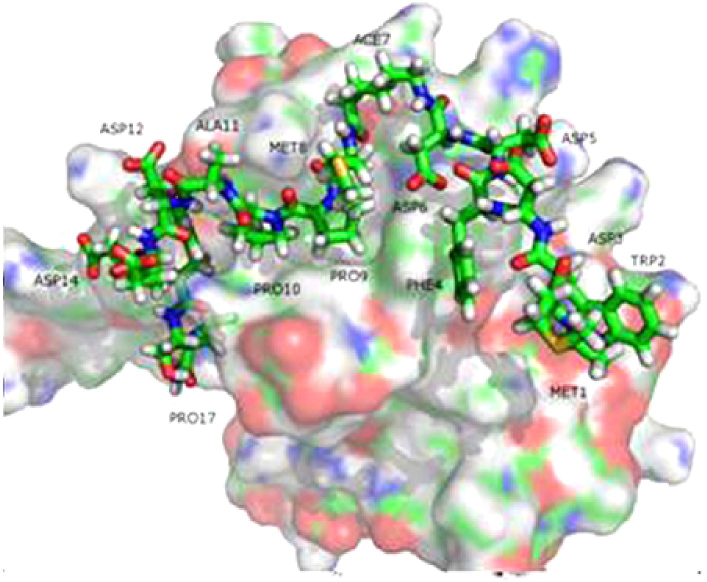


Figure 4.1: Annotated crystal structure of the CXCL8 ligand (surface) and its receptor peptide CXCR1 (stick) The color-coded representation of the receptor peptide left panel [oxygen (red), carbon (green), nitrogen (blue), hydrogen (white), and sulfur (yellow)] illustrates the close fit to the ligand protein.

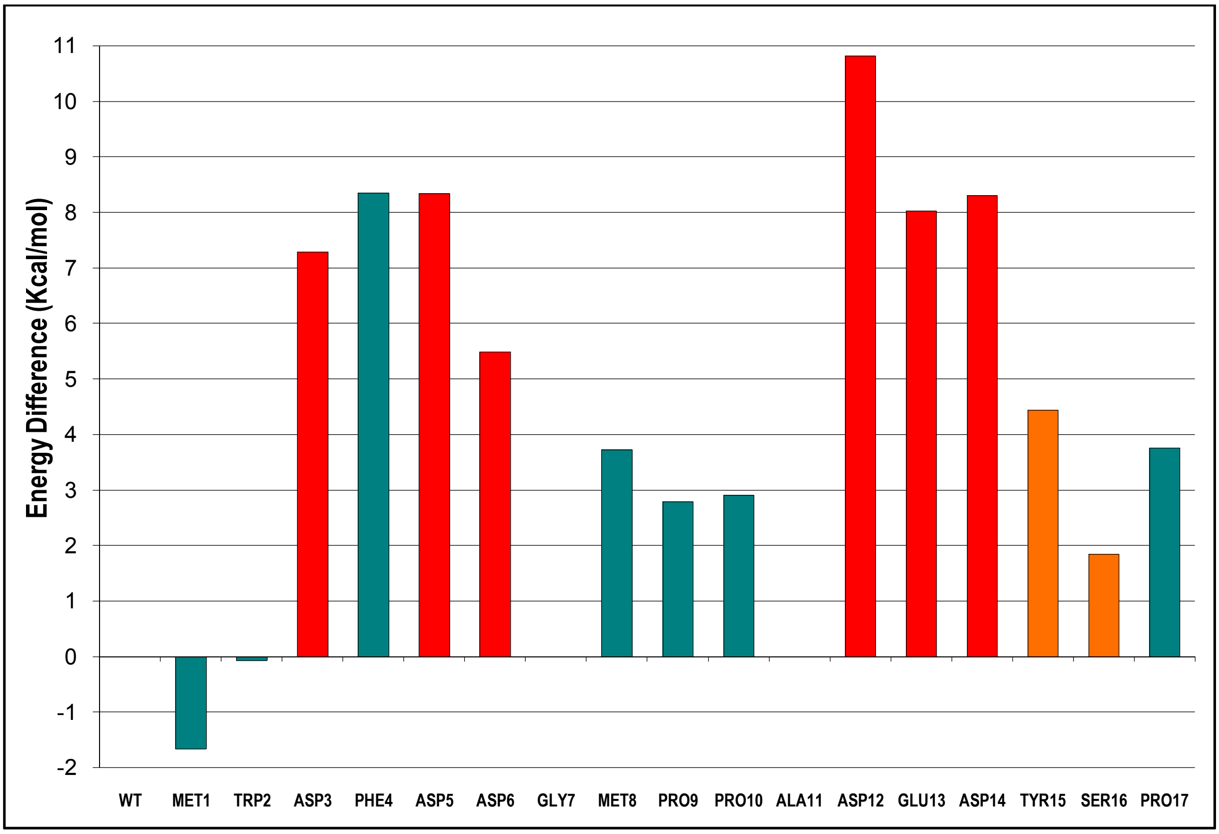


Figure 4.2: The histogram reports the energy differences (in Kcal/mol) for all 17 residues of the N-terminal peptide of CXCR1 in complex with chemokine CXCL8. Colours of the columns identify different kinds of amino acids: Red for AA with a polar acidic side chain; green for AA with a non polar side chain and orange for AA with a polar side chain.

We applied this analysis to a total of 17 residues from the N-terminal of CXCR1 receptor peptide mimic, which were individually mutated to alanine. The results are summarized in figure 4.2. As expected, mutations in the first two residues, which hardly interact with the chemokine, have little effect on the binding energy. The strongest effect arises from mutations of charged residues to apolar alanine, which indicates strong electrostatic components for the stabilization of this complex, followed by the polar amino acids. Not surprisingly, the substitution of non-polar amino acids with alanine results in smaller changes in the stabilization energy in general. In total, a significant jump of energy was observed for 7 of the 17 N-terminal amino acids of CXCR1. These findings are in agreement with experimental observations, which found strong effects for mutations on Pro9, Tyr15, Pro17, Asp12, Glu13, and Asp14 (Skelton, Quan et al. 1999) and substitution of Asp3 (Baggiolini and Moser 1997).

4.4. Elucidating specificity for ERBIN and its receptor ERBB2

In the second part of the study we investigated the ERBIN/ERBB2 complex, which is involved in the sorting of tyrosine kinase ERBB2 to the basolateral membrane of epithelial cells (Jaulin-Bastard, Saito et al. 2001; Dillon, Creer et al. 2002). ERBIN, binds to ERBB2 via its PDZ domain, a peptide sequence which acts as a dominant basolateral targeting signal in epithelial cells (Borg, Marchetto et al. 2000; Jaulin-Bastard, Saito et al. 2001). The ERBB2 C-terminus contains two motifs corresponding to the consensus sequence of tyrosine-based basolateral targeting signals, *N*-*P*-*X*-*Y* and *Y*-*X*-*X*, where represents a hydrophobic residue. Dillon et al. (Dillon, Creer et al. 2002) mutated the critical tyrosine residue to alanine and confirmed that the mutation of tyrosine causes a significant redirection of the mutant protein to the apical membrane.

Since the structure of the ERBIN-PDZ domain bound to the C-terminal tail of the ERBB2 receptor has been elucidated (PDB code 1MFG), we chose this protein complex as another model for a computational alanine screen (Fig. 4.3). We have applied the simulation protocols described above to ERBIN in complex with the cytosolic domain of ERBB2, a receptor tyrosine kinase present on the basolateral membrane of polarized epithelia with important functions in organ development and tumorigenesis.

Of the 9 mutated amino acid residues only the alanine substitution of Tyr2 corresponding to tyrosine Tyr1248 of the full receptor protein sequence leads to a significant change in interaction energy. This agrees well with experimental observations and an analysis of the crystal structure in which the phenyl ring of Tyr1248 is accommodated in a pocket formed by Ser1296, Gly1301, Asn1304, and Pro1305 stabilized by hydrophobic effects and hydrogen bonds to asparagine Asn1304 via ordered water molecules. These stabilizing interactions are lost upon alanine substitution leading to the observed change in interaction energy. Phosphorylation of Tyr1248 similarly disturbs the interaction of ERBB2 and the ERBIN-PDZ domain which is likely to be the mechanism by which phosphorylation modulates in ERBB2 signaling (Birrane, Chung et al. 2003). Considering the decomposition of the energy contributions (Table 4.II) we find a pattern mirroring the results of the previous complex, where essentially all of the energy change is either due to electrostatic or solvent contributions.

Residue No.	Aminoacid	Solvation	Side chain electrostatics	Main chain electrostatics	Backbone torsion	Lennard-Jones	Total
1	Glu	0,53	-0,11	-0,05	0,05	0,12	0,54
2	Tyr	**5,42**	-0,23	-0,05	0,05	0,66	**5,85**
3	Leu	0,77	0,1	-0,06	0,05	0,28	1,14
4	Gly	0,53	0,1	-0,05	0,05	0,1	0,73
5	Leu	3,21	-0,36	-0,05	0,05	0,28	3,13
6	Asp	-0,46	0,44	-0,05	0,05	0,27	0,25
7	Val	2,72	0,1	-0,05	0,05	0,17	2,99
8	Pro	0,54	0,13	-0,12	0,05	0,15	0,75
9	Val	2,42	0,1	-0,06	0,05	0,48	2,99

Table 4.II: Energy contributions (in Kcal/mol) of alanine substitutions of individual amino acids in the N-terminal peptide of ERBIN. The total energy is split into contributions by solvation effects, electrostatics side and main chain, backbone torsion, and Lennard-Jones potential. Large changes in interaction energy can be attributed to either solvation effects in case of polar and nonpolar residues or to electrostatic effects for charged side chains. Bold typeface indicates stronger interactions.

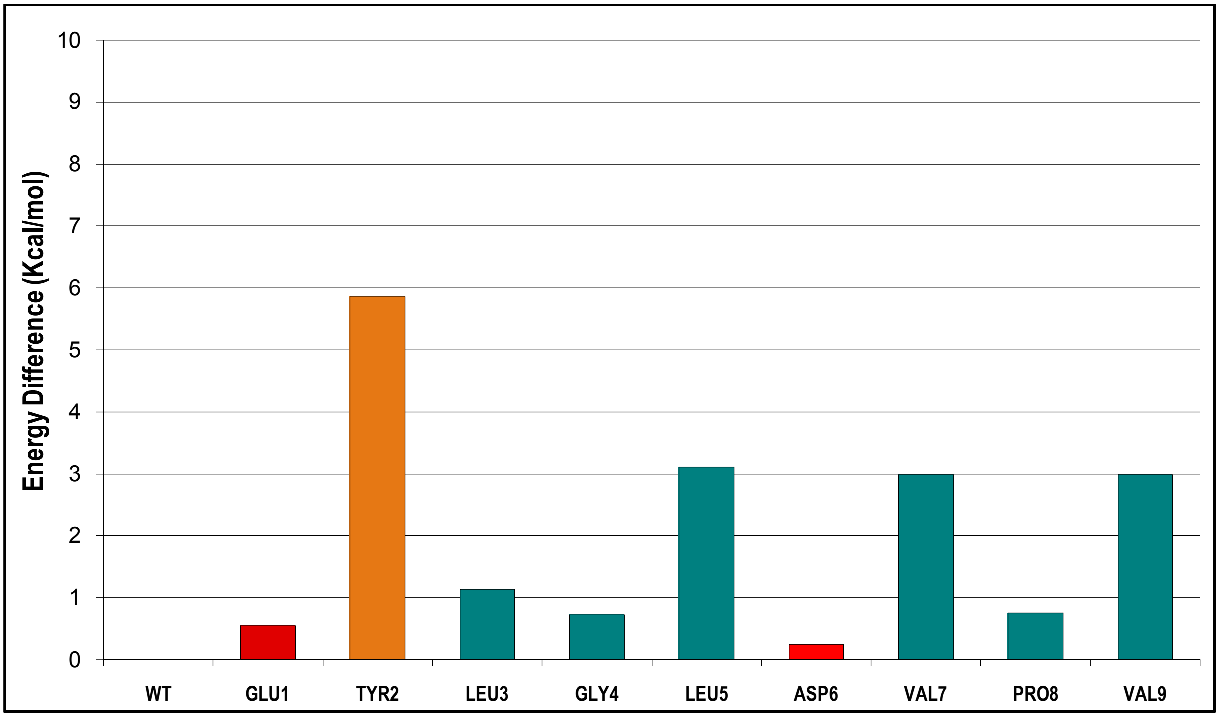


Figure 4.3: The histogram reports the energy differences (in Kcal/mol) for all 9 residues of the receptor peptide of ERBIN in complex with ERBB2. Colours of the columns identify different kinds of amino acids: Red for AA with a polar acidic side chain; green for AA with a non polar side chain and orange for AA with a polar side chain.

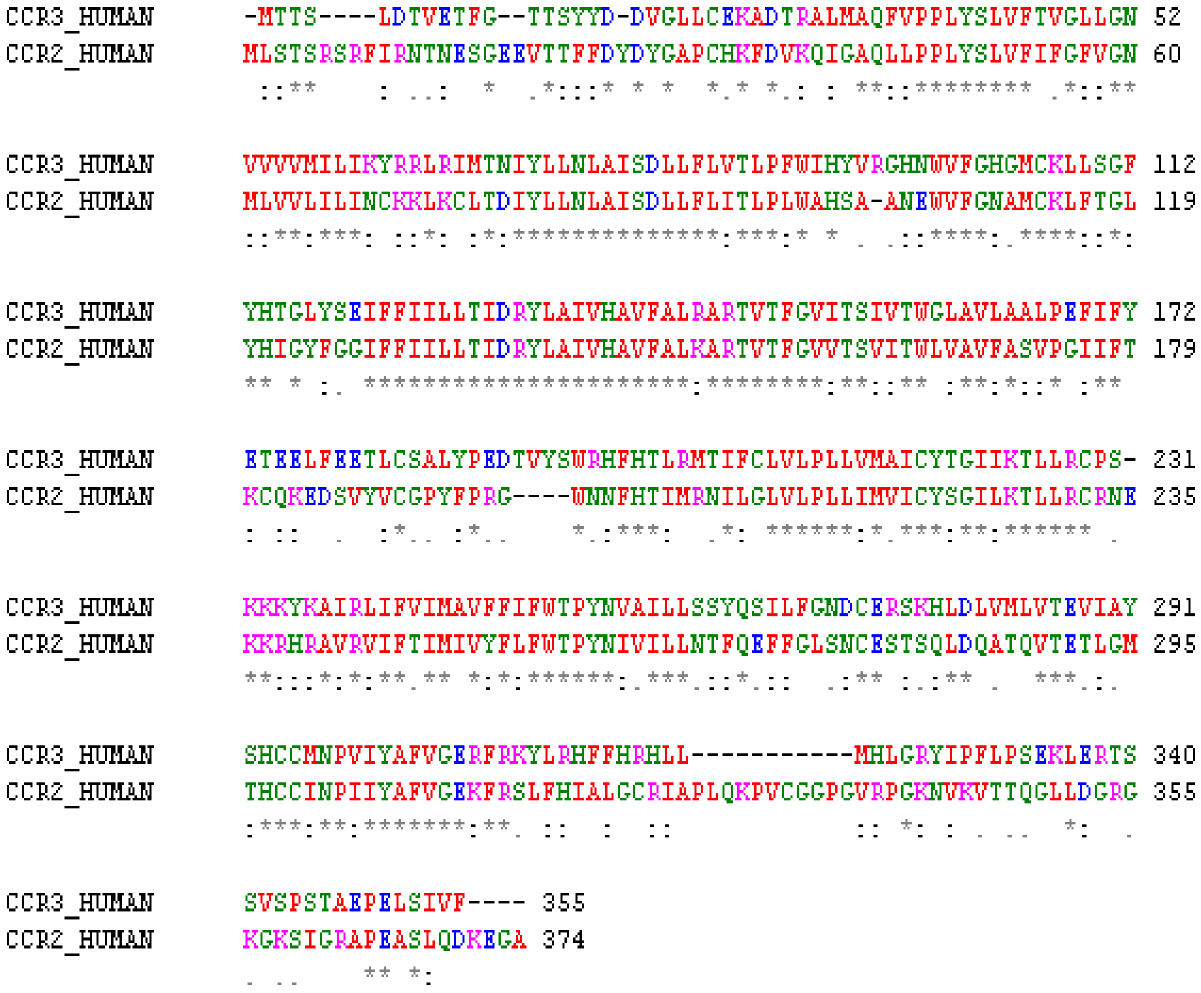
On the basis of the structure of the peptide ligand and parts of the receptor domain we were able to identify the most important interaction hot spot in agreement with experimental data (Dillon, Creer et al. 2002; Fan, Wong et al. 2005).

4.5. Elucidating specificity for eotaxin and its receptor CCR3

We extended this work to predict the effect of all possible mutations on the previously identified hotspots of the extracellular domain of receptor CCR3 that are crucial for eotaxin binding. The CC chemokine eotaxin plays a predominant role in eosinophil trafficking *in vivo* by specifically activating the chemokine receptor CCR3. CCR3 is a seven-helix transmembrane protein, and it is suggested that the binding site consists of a complex of several domains (Efremov, Truong et al. 1999; Ma, Bryce et al. 2002). This β chemokine receptor can be found on eosinophils, subsets of T_H2 cells, basophils, mast cells, neural tissue, and some epithelia cells (Tang and Powell 2001; Humbles, Lu et al. 2002). The first four cells types are crucial in allergy response (Humbles, Lu et al. 2002).

CCR3 is expressed on these cells why activation of CCR3 by eotaxin is believed to play a role in allergic diseases (e.g. asthma, dermatitis, and sinusitis) (Nickel, Beck et al. 1999) and in responses to parasitic infections (Jose, Adcock et al. 1994; Baggiolini 1996; Rankin, Conroy et al. 2000). CCR3 is specific for eotaxin and its analogs eotaxin-2/MIPF2 (Forssmann, Uguccioni et al. 1997) and eotaxin-3 (Kitaura, Suzuki et al. 1999) although CCR3 can also be activated by the chemokines RANTES, MCP-2, MCP-3, and HCC-2, which have lower receptor specificity (Ponath, Qin et al. 1996; Heath, Qin et al. 1997).

Elucidating the full binding motifs and locating their position on a three-dimensional model of the receptor CCR3 may help reduce the experimental effort to investigate chemokine receptor recognition and activation, but presently no crystal structure for these receptors is available. For this reason we have initiated our study by constructing an homology model for the CCR3 receptor. An initial template selection was made based on homology to the crystal structure of bovine rhodopsin (PDB code 1F88) (Palczewski, Kumasaka et al. 2000) and existing homology models of CCR5 and CCR2 (Shi, Liu et al. 2002; Liu, Hwangbo et al. 2004), the template selection was made using the 3DJURY (Ginalski, Elofsson et al. 2003), PHYRE (Kelley and Sternberg 2009). Based on these data, a model for Eotaxin receptor CCR3 was constructed on the region from 1 to 355 using as a template CCR2 chemokine receptor (PDB code 1KAD) (Singh and Somvanshi 2009). The two receptors CCR3 and CCR2 are very similar in structure and function so we used ClustalW (Thompson, Higgins et al. 1994) to generate an alignment between CCR3 and CCR2 (Fig.4.4). We used MOE to generate a final model on the basis of the alignment (Fig.4.5).



```
CCR3_HUMAN      -MTTS----LDTVETFG--TTSYYD-DVGLLCEKADTRALMAQFVPPLYSLVFTVGLLGN 52
CCR2_HUMAN      MLSTSRSRFIRNTNESGEEVTTFFDYDYGAPCHKFDVKQIGAQLLPPLYSLVFIFGFVGN 60
                 ::**    : ..:  *  .*:::* * *  *.* *.: : **::******** .*::**

CCR3_HUMAN      VVVVMILIKYRRLRIMTNIYLLNLAISDLLFLVTLPFWIHYVRGHNWVFGHGMCKLLSGF 112
CCR2_HUMAN      MLVVLILINCKKLKCLTDIYLLNLAISDLLFLITLPLWAHSA-ANEWVFGNAMCKLFTGL 119
                ::**:***: ::*: :*:**************:***:* * . .::****:.****::*:

CCR3_HUMAN      YHTGLYSEIFFIILLTIDRYLAIVHAVFALRARTVTFGVITSIVTWGLAVLAALPEFIFY 172
CCR2_HUMAN      YHIGYFGGIFFIILLTIDRYLAIVHAVFALKARTVTFGVVTSVITWLVAVFASVPGIIFT 179
                ** * :. **********************:********:**::** :**:*::* :**

CCR3_HUMAN      ETEELFEETLCSALYPEDTVYSWRHFHTLRMTIFCLVLPLLVMAICYTGIIKTLLRCPS- 231
CCR2_HUMAN      KCQKEDSVYVCGPYFPRG----WNNFHTIMRNILGLVLPLLIMVICYSGILKTLLRCRNE 235
                : ::  .  :*.. :*..    *.:***:  .*: ******:*.***:**:****** .

CCR3_HUMAN      KKKYKAIRLIFVIMAVFFIFWTPYNVAILLSSYQSILFGNDCERSKHLDLVMLVTEVIAY 291
CCR2_HUMAN      KKRHRAVRVIFTIMIVYFLFWTPYNIVILLNTFQEFFGLSNCESTSQLDQATQVTETLGM 295
                **:::*:*:**.** *:*:******:.***.::*.::  .:** :.:** .  ***.:.

CCR3_HUMAN      SHCCMNPVIYAFVGERFRKYLRHFFHRHLL-----------MHLGRYIPFLPSEKLERTS 340
CCR2_HUMAN      THCCINPIIYAFVGEKFRSLFHIALGCRIAPLQKPVCGGPGVRPGKNVKVTTQGLLDGRG 355
                :***:**:*******:**. ::  :  ::            :: *: : . ..  *:  .

CCR3_HUMAN      SVSPSTAEPELSIVF---- 355
CCR2_HUMAN      KGKSIGRAPEASLQDKEGA 374
                . ..    ** *:
```

Figure 4.4: Alignment between CCR3 and CCR2, performed using ClustalW.
Color corresponds: red, AVFPMILW Small (small+ hydrophobic (incl.aromatic -Y)); Blue, DE Acidic; Magenta, RHK Basic;Green, STYHCNGQ Hydroxyl + Amine + Basic – Q.

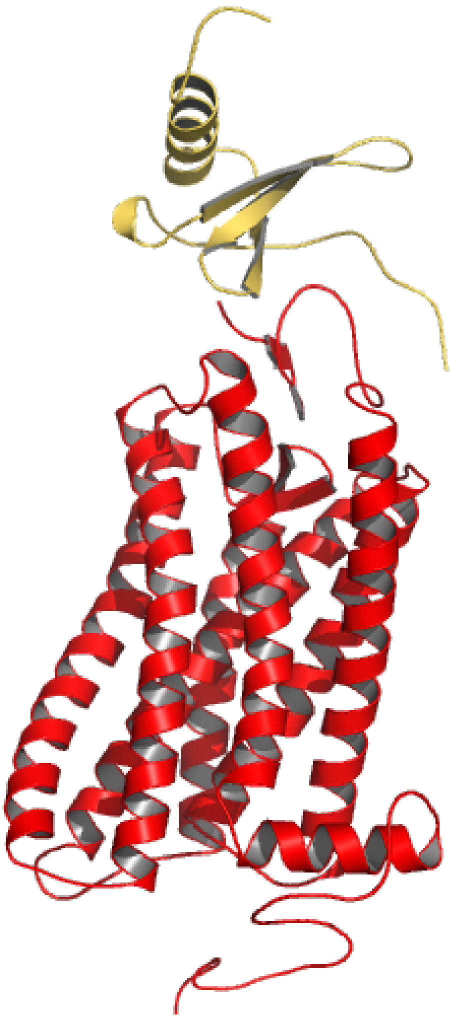


Figure 4.5: Model for complex of the CCR3 receptor (red) with eotaxin ligand (yellow).

We applied this analysis to a total of 35 residues from the N-terminal of CCR3 receptor, which were individually mutated all the possible 20 amino acids. The results are summarized in figure 4.6. The result matrix colored each mutation on a red-green gradient depending on the lower (red) or stronger (green) binding energy interaction compared with the native structure. Mostly all residues showed little change on the binding energy compared to the WT. 15 residues showed a stronger binding energy interaction compared to the native structure, in particular the amino acids ILE and LEU for the first residue of the N-terminal of the receptor. To demonstrate our findings peptide sequences from chemokine receptors containing the most promising hot spot amino acids from the list will be synthetized and tested in a set of binding experiments by the group of Dr. Katja Schmitz.

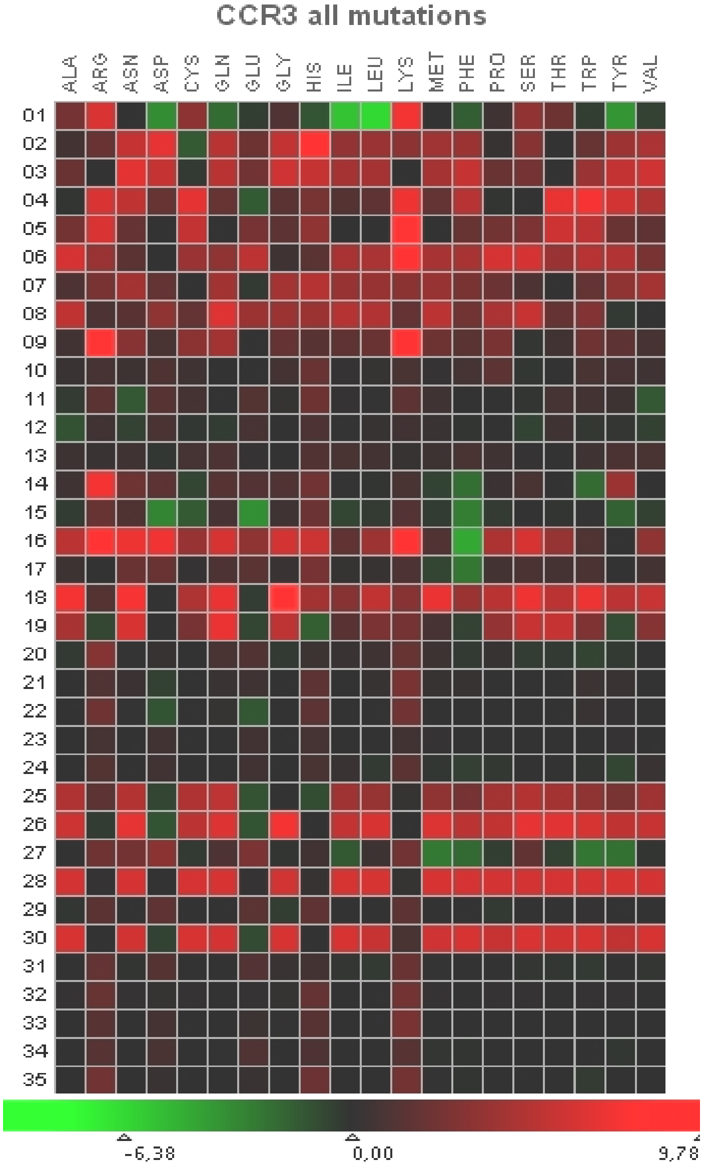


Figure 4.6: Results of all the possible mutations of the Receptor CCR3, where green spots mean stronger interaction compared to the original structure and red spots mean fewer interaction compared to the original structure.
X-axis on top shows all the 20 amino acid; X-axis on bottom shows the energy values (Kcal/mol); Y-axis shows all the 35 residue mutated of the CCR3 receptor.

4.6. Summary

Because protein-protein interactions are so important in regulating biological pathways and increasing information about key molecules involved in various different regulatory pathways continuous to become available, modulation of protein-protein interactions has become an important goal to affect biological function (Wells and McClendon 2007).

The experimental study of hot spots in protein-protein interfaces has generated a large amount of structural and functional data, which has been compiled in hot spot databases (Thorn and Bogan 2001; Fischer, Arunachalam et al. 2003). Collections of experimental data have, in turn, been used to develop and test computational methods for the detection of hot spot residues (Guerois, Nielsen et al. 2002; Kortemme and Baker 2002; Darnell, Page et al. 2007; Ofran and Rost 2007) and to derive common characteristics of hot spots, such as their clustering in densely packed regions with an abundance of conserved polar residues and cooperativity of hot spot regions (Bogan and Thorn 1998; Hu, Ma et al. 2000; Ma, Elkayam et al. 2003; Keskin, Ma et al. 2005).

Different approaches have been used to predict hot spots: in a physical and knowledge-based approach neural networks have been trained to recognize specific features of interface residues leading to the KFC model for the rapid prediction of hot spot residues from structural data (Darnell, Page et al. 2007; Darnell, LeGault et al. 2008).

Training of a network on sequence environment, environmental trace, and accessible surface area permits the prediction of hot spots with high positive accuracy from sequence data alone (Ofran and Rost 2007). Although methods based on large amounts of empirical data provide reliable predictions with a minimum of computational resources, the role of individual features in the decision making of the neural network may not be straightforward to interpret. Clustering of conserved residues within “hot regions” and their preorganization in the unbound state were found to be characteristic of hot spots in studies employing molecular dynamics for conformational analysis of individual residues after multiple structure alignment of complexes obtained from PDB (Ma, Elkayam et al. 2003; Keskin, Ma et al. 2005; Yogurtcu, Erdemli et al. 2008).

Modeling protein-protein interactions can aid our understanding of fundamental biological processes and help to design lead structures in drug development. Free-energy methods, as the one pursued here, are similar in spirit to MM-GBSA, but model conformational change solvent accessible surface area SASA based solvation model of a given backbone conformation. Our approach has been used to predict the native conformation of a number of small proteins up to 60 amino acids in size, which documents its use in describing intra-molecular interactions, but was not previously applied to problems of protein-protein interactions.

With the studies reported in this chapter I have first done an alanine scanning of the first 17 residues of the N-terminal of the chemokine receptor CXCR1 and for the 9 residues of the C-terminal of the ERBB2 receptor.

For CXCR1 a significant change in energy was observed for 7 of the 17 N-terminal amino acids and for ERBB2 only one amino acid led to a significant change in interaction energy.

To complete the investigation 35 residues from the N-terminal of CCR3 receptor, were individually mutated with all the possible 20 amino acids. Some residues showed a stronger binding energy interaction compared to the native structure.

5. Protein Ligand Interactions

5.1. Introduction

Protein-ligand interactions are essential for almost all biological processes. For many proteins involved in diseases the biophysical mechanisms that activate potential binding partners to associate remains poorly understood. New developments in protein-ligand interactions are the studies in chemogenomics which are among the key challenges for drug discovery. Chemogenomics use the comprehensive genomic data available after the elucidation of the human genome and others in order to identify effective new medicines (Agrafiotis, Lobanov et al. 2002; Zheng and Chan 2002; Klabunde 2007). For example numerous chemogenomic approaches apply the classification of target families (such as ion channels, kinases, GPCRs) or protein subfamilies (such as purinergic GPCRs) without taking into account similarities of the assumed ligand-binding sites (Rognan 2007).

In collaboration with the experimental groups of Dr. Katja Schmitz at KIT (Karlsruhe Institute of Technology) (section 5.3), Prof. Stefan Bräse (KIT) (section 5.4) and Dr. Pérez-Sánchez with Prof. J. Corral at the University of Murcia (section 5.5), we have investigated the interactions of small molecule binding protein receptors and performed virtual screening for drug discovery. Specifically we have first performed a binding study of small molecules synthetized by the group of Dr. Katja Schmitz with the chemokine receptor CXCR1 to further extend the work reported in the previous chapter (see section 4.1 for general information on chemokines). The aims of this study were to identify possible binding sites of the molecules with IL8 and to rank the small-molecule IL8 interactions for the different candidate binding sites.

In a second study we have constructed an homology model for CB1 and CB2 receptors and performed docking simulations for a set of compounds that have been investigated experimentally in the group of Prof. Stefan Bräse and Prof. Christa Müller. To guide further synthesis efforts we studied a set of derivatives of the coumarine scaffold, which was functionalized at a specific position with aliphatic sidegroups in order to increase the affinity. The primary goal of our simulations was to understand changes in the binding mode and the binding energy and to develop structurally related novel compounds with a better affinity.

Finally in collaboration with the University of Murcia (Spain), we developed a structure-based model for AT-II regulation to develop new anticoagulation drugs. We performed docking simulations for the complex antithrombin/heparin, which were followed by a large-scale *in silico* screen comprising over 800.000 ligands that resulted in a new promising compound to regulate blood coagulation in humans. Before reporting in detail in these studies in sections 5.5.1-5.5.2 I review the methods used in section 5.2. and conclude with a general overview in section 5.6.

5.2. Methods

FlexScreen (Fischer, Basili et al. 2007) performs fully automated *in silico* screening of a large 3D database of ligands against a structurally resolved protein receptor. Each ligand of the database is docked against the receptor with the stochastic tunneling method (Wenzel and Hamacher 1999) using an all-atom representation of both ligand and receptor. The scoring function contains a sum of the van der Waals, electrostatic, hydrogen-bond and solvation energies. The VdW and hydrogen-bond parameters are taken from OPLSAA (Jorgensen and N.A. 1997) and AutoDock (Goodsell, Morris et al. 1996), respectively, the partial charges of the receptors are computed with MOE, and the atomic solvation parameters where solvation energy is described as a sum of energies for the individual atoms that are assumed to be proportional to the solvent accessible surface area and atomic solvation parameter, ASP. All atoms are divided into two groups: hydrophobic and hydrophilic effects, so that only two ASPs have to be optimized.

The docking simulations use a cascadic approach: the total number of simulation steps is divided into several partitions of similar computational effort. In the first partition 100 simulations with 7500 computational steps, in the second partition 5 simulations with 30000 computational steps and in the third partition 2 simulations with 75000 computational steps are performed. In partition 2 and 3 only the best energetic trajectories of the former partition are continued. To avoid in partition 2 a simulation of nearly identical conformations, we divide the final 100 conformations in stage 1 into three different clusters: cluster 1 includes the best-scoring conformation and all others with a similar binding pose (RMSD less than 0.8 Å), cluster 2 includes the best-scoring ligand outside cluster 1 and all conformations with a RMSD of 0.8 Å and cluster three contains all other conformations. We start stage 2 with the two top-scorers of clusters 1 and 2 and the top-scorer of cluster 3. For partition 1 we use three different starting conformations which are randomly selected for each of the 100 simulations: a relaxed conformation, a conformation with largest atom-to-atom distance and a conformation with smallest radius of gyration.

The cascadic approach balances diversity and computational effort and invests the largest computational effort into the most promising candidates. The difference in the scoring function for the final results of partition 3 can be used to estimate the error and may indicate if the binding mode is unique. For statistical reasons we repeated each docking run ten times. To limit the effect of outliers, we use the median of the 10 results for correlation analysis.

Each of the simulations in the cascade employs the stochastic tunneling method (STUN) as the global optimization technique (Wenzel and Hamacher 1999). This method allows an all-atom minimization process to "tunnel" through forbidden regions of the potential energy surface, which is subject to a non-linear energy transformation:

$$E_{STUN} = ln(x + \sqrt{x^2 + 1})$$

Here, $x = \gamma\ (E(r) - E_0)$. E_0 is the lowest minimum encountered by the dynamical process so far, and γ is a problem-dependent parameter, which controls the steepness of the transformation.

Scoring performance of the FlexScreen approach was benchmarked by using of the ASTEX database based on over 80 target proteins of different classes with available ligand-bound X-ray crystal structures (Fischer, Basili et al. 2007). The ASTEX/CCDC receptor-ligand database was also used to study the accurancy of the binding modes performance where over 80% of the ligands dock within 2 Å of the experimental binding mode and for 60 complexes the docking protocol locates the correct binding mode in all of ten independent simulations. Selectivity was investigated in an analysis of 12 dataset of the DUD database (Kokh and Wenzel 2008). For each protein the database includes: 1) crystal structures of the receptor and its native ligand; 2) a set of the annotated ligands that should in principle dock well (15-450 molecules); 3) a set of the decoys (about 36 molecules for each annotated ligand) that resemble the particular ligand in physical properties, but differ from the ligand topologically, so that they are likely to be nonbinders (Huang, Shoichet et al. 2006).

Investigation on the enrichment rates of rigid-, soft- and flexible-receptor models was also performed for 12 diverse receptors using libraries containing up to 13000 molecules (Kokh and Wenzel 2008). For the rigid-receptor model, a high enrichment ($EF_1 > 20$) was observed, but only for four target proteins. For the soft-receptor model improved docking rates were observed, but showed a reduced enrichment rates. Finally for the flexible side-chain model with flexible dihedral angles for up to 12 amino acids (3−8 flexible side chains), showed an increasement of both binding propensity and enrichment rates.

5.3. Small Molecule Ligands for Binding Chemokines

A promising avenue for drug discovery would be the rational design of peptides or small molecules that bind strongly to the chemokine to prevent the chemokine from binding to its receptor. After the complete scanning mutation for the two chemokine receptors CXCR1 and CCR3 reported in section 4.5, we performed in collaboration with Dr. Katja Schmitz a study on small molecules that could bind to the chemokine receptor CXCR1. The binding energy of a set of 19 small molecules, which were active in a chip assay against chemokine interleukin-8, was investigated by receptor-ligand binding simulations. From the assay data it remained uncertain, whether the molecules would bind specifically or non-specifically to IL8. Only molecules binding at or near the receptor binding site IL8 are likely to interfere with chemokine receptor binding.

The aims of this study were therefore to identify possible binding sites of the molecules with IL8 and to rank the small-molecule-IL8 interactions for the different candidate binding sites. I have used Qsitefinder (a method of ligand binding site prediction) (Laurie and Jackson 2005) to identify 10 possible binding sites (Fig. 5.1), which cover essentially the whole protein surface, including the receptor binding site of IL8 (binding site 1 and 9, Fig. 5.2). Each ligand was docked independently at each of the 10 docking site.

The energies of the lowest energy binding pose for each ligand and site are summarized in Table 5.I and figure 5.3. The data demonstrate that there are a few sites, where the molecules may weakly bind. Molecule 3, 4, 17 and 19 have the best energy at binding sites 7, 8, 9, respectively, while binding site 6 shows a very low binding energy compared to the others for all ligands. Binding site 9 (Fig. 5.2) is near the position where interleukin-8 interacts with his receptor. In view of these results binding site 9 poses a good candidate to be tested experimentally with the 4 small molecules found to interact strongest with this binding site.

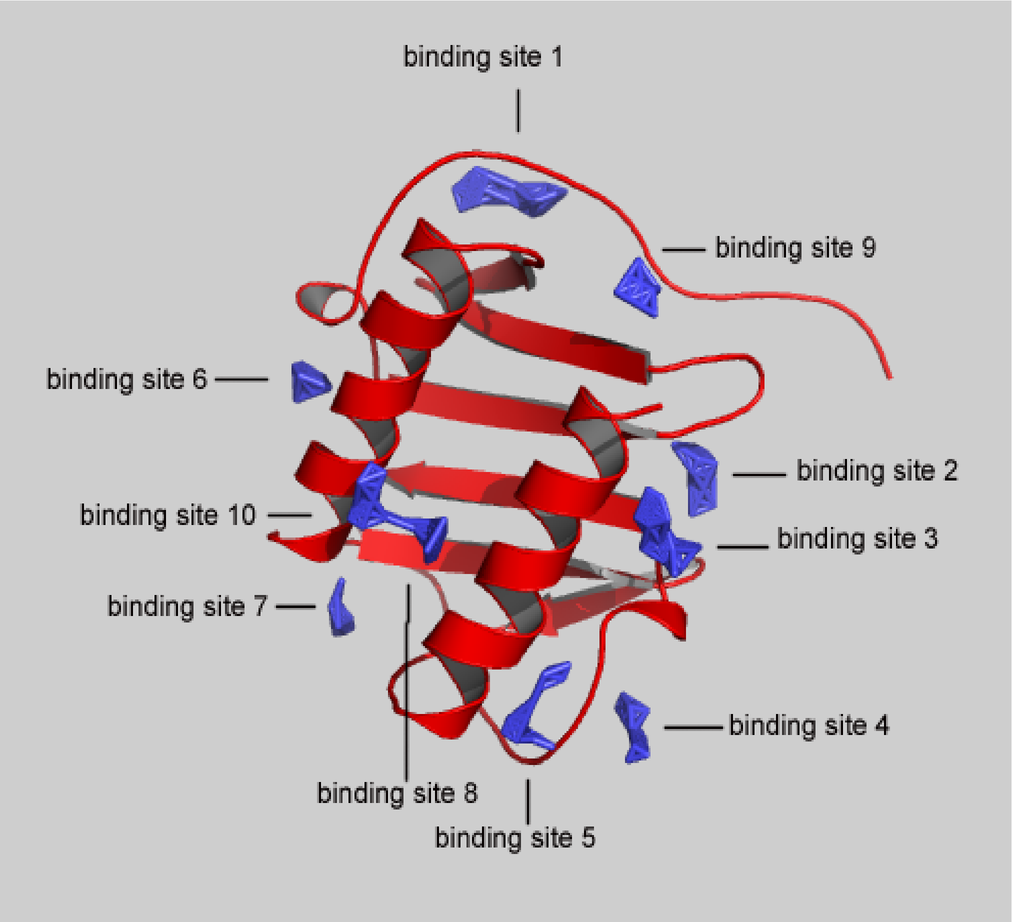


Figure 5.1: Illustration of the 10 binding sites predicted for Inteleukin-8 using Q-SiteFinder.

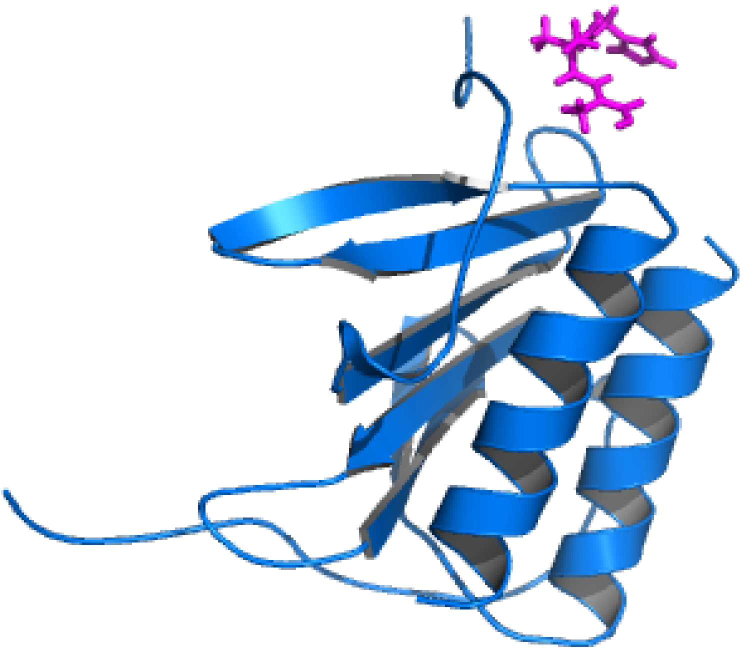


Figure 5.2: Binding site 9 between CXCR1 (blue) and one of the 19 small molecules (magenta) after prediction with Qsitefinder.

PEPTIDE RECEPTORS	BINDING ENERGY 1	BINDING ENERGY 2	BINDING ENERGY 3	BINDING ENERGY 4	BINDING ENERGY 5	BINDING ENERGY 6	BINDING ENERGY 7	BINDING ENERGY 8	BINDING ENERGY 9	BINDING ENERGY 10
Molecule 1	-128,28	-136,91	-153,92	-113,84	-120,56	-97,58	-127,06	-149,29	-136,69	-150,85
Molecule 2	-224,35	-201,46	-235,15	-173,46	-172,00	355,21	-224,63	-225,29	-234,34	-223,77
Molecule 3	-255,12	-266,04	-282,73	-294,18	-267,36	0,00	-372,61	-387,64	-416,29	-326,21
Molecule 4	-242,4573	-226,11	-269,31	-177,58	-200,10	-143,81	-266,35	-272,59	-307,03	-277,36
Molecule 5	-126,45	-130,35	-135,24	-106,88	-110,75	-63,78	-141,49	-119,20	-140,22	-129,00
Molecule 6	-71,18	-80,82	-100,31	-73,53	-72,51	-73,43	-113,47	-89,52	-116,84	-89,58
Molecule 7	-85,36	-58,92	-86,22	-63,99	-63,73	-44,40	-97,89	-75,99	-83,08	-77,27
Molecule 8	-105,73	-102,77	-137,66	-101,04	-85,96	1897,69	-110,06	-117,11	-112,06	-117,84
Molecule 9	-138,13	-118,35	-156,16	-110,86	-104,78	35,00	-135,17	-144,52	-148,18	-134,07
Molecule 10	-143,93	-129,45	-150,25	-120,87	-116,51	178,81	-135,05	-107,81	-161,24	-174,27
Molecule 11	-90,43	-78,21	-100,52	-75,97	-80,16	-92,45	-89,44	-95,48	-104,12	-95,04
Molecule 12	-104,53	-109,19	-122,25	-114,25	-92,56	7308,75	-142,01	-141,23	-141,90	-152,29
Molecule 13	-83,66	-126,16	-105,59	-106,94	-89,48	1179,04	-142,86	-133,79	-130,63	-136,56
Molecule 14	-128,59	-146,23	-150,29	-110,22	-102,39	388,59	-127,87	-141,85	-138,02	-134,54
Molecule 15	-88,40	-79,54	-97,61	-73,90	-60,26	-65,35	-80,99	-79,09	-77,19	-79,96
Molecule 16	-96,88	-106,99	-108,26	-98,58	-81,55	-74,12	-109,31	-114,69	-125,93	-119,46
Molecule 17	-198,43	-216,60	-248,83	-188,75	-140,81	0,00	-258,49	-289,23	-288,31	-239,82
Molecule 18	-102,53	-78,71	-82,39	-82,31	-76,89	-31,88	-67,36	-90,30	-75,99	-90,32
Molecule 19	-239,28	-187,79	-237,06	-167,76	-175,02	-146,29	-229,68	-234,22	-244,15	-231,33

Table 5.I: Summary of the binding energies of the peptides receptor for every binding site (name of the compounds are in the APPENDIX).

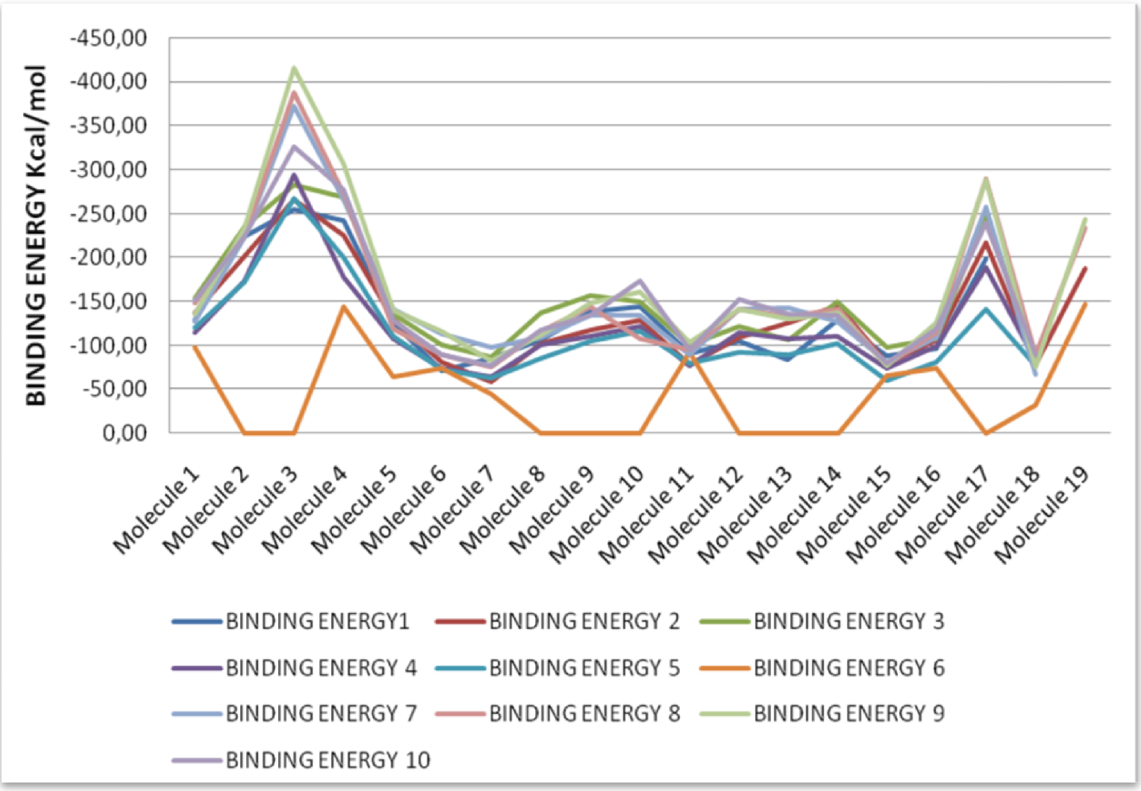


Figure 5.3: Histogram summary of all small molecules binding with the 10 differents binding site. Molecule 3, 4, 17 and 19 have the best energy at binding sites 7, 8, 9, respectively, while binding site 6 shows a very low binding energy compared to the others for all ligands.

5.4. Selective high-affinity inhibitors of cannabinoid receptors CB1/CB2

Cannabinoid receptors, which are members of the membrane-bound G protein-coupled receptor superfamily, are involved in neuroinflammatory and neurodegenerative disorders, such as Huntington and Alzheimer disease (Bisogno and Di Marzo 2010; Van Laere, Casteels et al. 2010). Cannabinoid receptors are activated by three major groups of ligands, mammalian endocannabinoids, plant and synthetic cannabinoids (e.g. THC from the cannabis plant). In collaboration with the group of Prof. Bräse at KIT and Prof. Müller at PharmaCenter Bonn, I have investigated the interaction between the Cannabinoid receptors 1/2 (CB1/CB2) with a novel set of coumarin derivatives. Rational improvement of known compounds to inhibit these receptors is complicated by the lack of available crystal structures, which is a general problem for GPCRs, which nevertheless constitute about 30% of all presently known drug targets. In addition such investigations are complicated by the fact, that selectivity between the receptor classes is desired, because, molecules can recognize two or more receptors and display a different affinity for them. I have therefore generated homology models for the two receptors (section 5.4.1) and then performed docking simulations for a set of small molecules that were experimentally validated (section 5.4.2). In particular for the CB1 receptor a good correlation between the experimental and theoretical results could be observed, comparison between the CB1 and CB2 receptor models also helped to rationalize the selectivity of some compounds, as discussed in section 5.4.3.

The primary goal of our simulations was to understand changes in the binding mode and the binding energy between the new compounds and structurally related compounds with a known affinity.

The compounds used were coumarin derivatives, where coumarin is specifically a benzopyrone found in many plants, and is used in the pharmaceutical industry as a precursor molecule in the synthesis of several synthetic anticoagulants similar to dicoumarol and some even more potent rodenticides, that work by the same anticoagulant mechanism.

5.4.1. Model Building and Validation

All-atom models for the CB1 and CB2 receptors were constructed using the crystal structure of bovine rhodopsin (PDB code 1U19) (Okada, Sugihara et al. 2004) as structural template, to which both CB-receptors have a strong similarity in sequence. We constructed a model on the region from 80 to 439 for CB1 and from 1 to 349 for CB2 (Fig. 5.4). Template selection was performed using PHYRE (Kelley and Sternberg 2009) using the default protocol and the alignment between the receptors and the template was assessed using ClustalW (Thompson, Higgins et al. 1994). On the basis of the resulting alignment ten different models were built using MOE of which the model with the lowest energy profile was chosen for this investigation. As the template used (bovine rhodopsin), the two receptors CB1/CB2 show the typical structure of the G-protein coupled receptor family. GPCRs are characterized by an extracellular N-terminus, followed by seven transmembrane (7-TM) α-helices (TM-1 to TM-7) connected by three intracellular and three extracellular loops, and finally an intracellular C-terminus. The seven transmembrane helices are forming a cavity within the plasma membrane that serves as ligand-binding domain.

For the docking simulations we used FlexScreen (Merlitz, Burghardt et al. 2003; Fischer, Basili et al. 2007) receptor-ligand docking software with a SASA based implicit solvation model (Lee and Richards 1971). All simulations were performed using the homology models described above. In this study we used two different protocols: in the automatic docking protocol: - each ligand was docked against the receptor with the stochastic tunneling method using an all-atom representation of both ligand and receptor using a cascadic docking protocol. Both ligand and receptor can change their conformation in the docking process. SASA, the accessible surface area (ASA) is the surface area of a protein that is accessible to a solvent - in the relaxation protocol we started from the known binding mode of one ligand, superimposed, as closely as possible to related ligands synthetically derived by altering one or more substituents from the original ligand and performed only one long relaxation simulation. At the end of each simulation a binding energy for the ligand is computed as the difference between the unbound and bound complex using the biophysical scoring function of FlexScreen.

Site-specific mutation studies (Shim, Welsh et al. 2003; Tuccinardi, Ferrarini et al. 2006; Tuccinardi, Cascio et al. 2007) on the rhodopsin subfamily of receptors, including CB1/CB2, suggest that many ligands bind CB1/CB2 within the TM core region in the crevice formed by TM3, TM4, TM5, and TM6. Initial docking studies using AM281, CP55940 and win55, also investigated in Behrenswerth et al. (Behrenswerth, Volz et al. 2009), demonstrated that these reference ligands are well accommodated within the transmembrane region.

In the binding poses we observe a close geometric proximity of these ligands with amino acids PHE191(TM3), LYS192(TM3), VAL196(TM3), THR197(TM3), PHE200(TM3), TRP241 (TM4), ALA244(TM4), PHE278(TM5), TRP279(TM5), ARG340(TM6), CYS355(TM6), TRP356(TM6), LEU359(TM6), LEU360(TM6), MET363(TM6), CYS386(TM7), LEU387(TM7), LEU388(TM7), as illustrated in the pose of MAK15 in figure 5.6. This is in agreement with studies that reported alanine substitution of LYS192 to results in a significant loss of affinity for CP55940 (Tuccinardi, Ferrarini et al. 2006). On the other hand, mutation of PHE191, TRP279 and TRP356 to ALA showed a loss of affinity for win55. In addition, it was suggested that TRP356 might be important for win55 binding in CB1.

When we analyze the binding modes of AM281 as a reference ligand, we find a large hydrophobic pocket in vicinity of the docking pose that is formed by the sidechains (ALA198, CYS264, TRP279, TRP356, LEU359, MET363, PHE379, CYS386) (Fig. 5.5A), but which is not occupied by the ligand. Modifications of the ligand that fill this hydrophobic pocket with apolar substituents should dramatically improve the binding energy by exploitation of the hydrophobic effect. This observation is supported by that fact that in the binding pose of CP55940 we find its aliphatic sidechain to fill exactly this pocket (Fig. 5.5B).

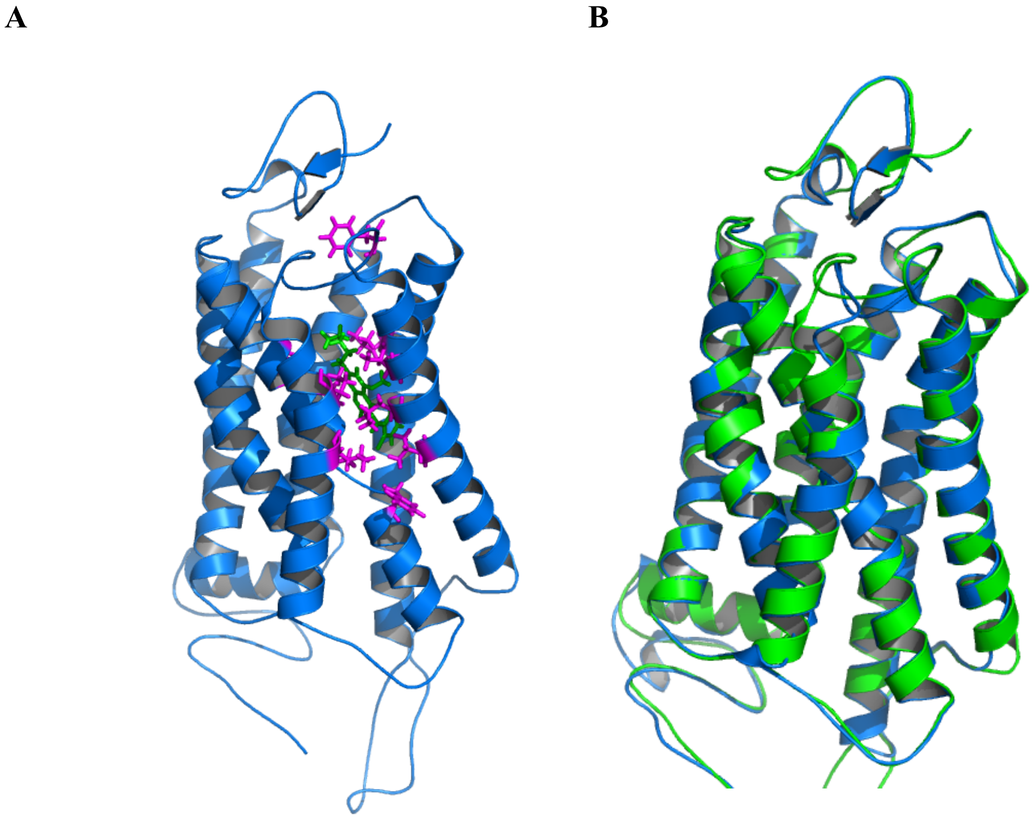


Figure 5.4: (A) Model for complex of the CB1 receptor with MAK15, illustrating the binding pocket and key interactions residues in the vicinity of the binding pocket. The ligand is stabilized by interactions with the marked residues (see text). (B) Overlay of the CB1(green) and CB2 (blue) receptor where a number of amino acids with relatively small sidechains (CB1: ALA200, ASP268, THR285, VAL293) are replaced by amino acids with relatively large sidechains (CB2: MET, GLU, ILE, ILE).

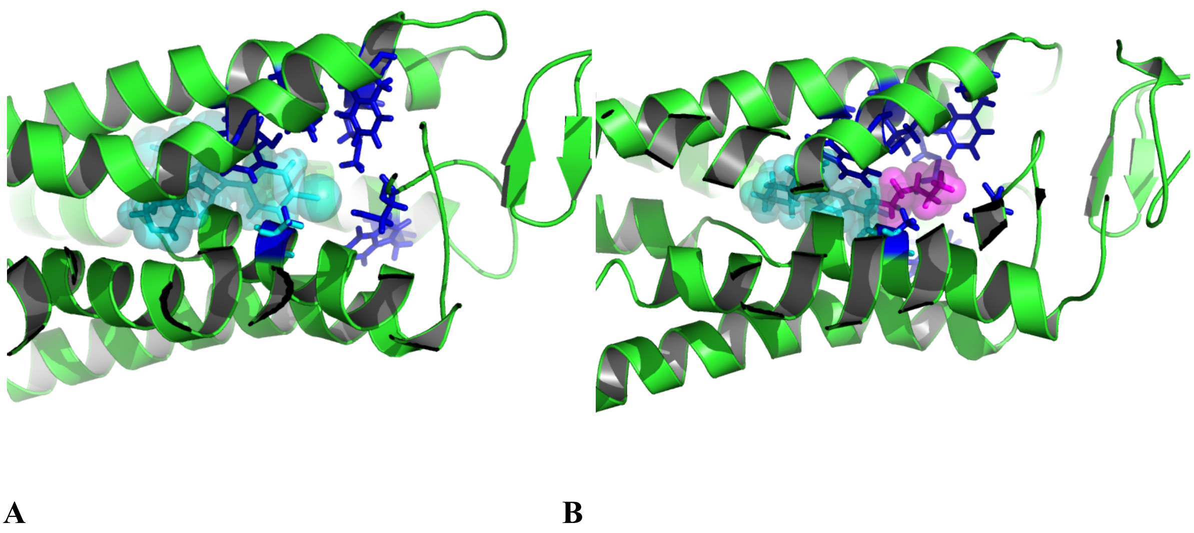


Figure 5.5: Hydrophobic pocket formed by ALA198, CYS264, TRP279, TRP356, LEU359, MET363, PHE379, CYS386 (in blue) of the CB1 receptor interacting with the reference ligands (A) AM281 and (B) CP55,940. We note that the aliphatic sidechain in CP55,940 (shown in magenta), occupies the hydrophobic pocket thereby improving the affinity.

5.4.2. Design of novel ligands

We have therefore docked a proposed set of new ligands with various aliphatic substitutions into the same binding pocket using the automated docking protocol and observed that the new ligands assume binding poses where the coumarine and aliphatic parts of the ligands overlap. Prediction of absolute affinities is difficult with state-of-the-art techniques even when a high quality crystal structure of at least one representative complex is available. Such a prediction is thus unlikely to succeed when only an homology model of the receptor is available. We therefore used the relaxation protocol (see methods of this section) in simulations of all 39 coumarine derivatives investigated by Behrenswerth et al., each of which showed a lower affinity than the reference compounds (see Table 5.II). We then applied the same simulation protocol to the new compounds, which were functionalized at position R7 with aliphatic sidegroups.

To rationalize the relative binding energies/affinities we concentrate on those compounds that are related by a single substitution of a side-group: ligands 25, 27, 24, 30 each differ from MAK10, MAK15, MAK17 and NAV478, respectively, by the aliphatic substitution at the R7 position, respectively. In addition MAK11 and MAK13 are similar to ligand 27 and 25, but here the position of the chlorine and the methyl group are also changed.

Comparison of the binding poses between the new compounds and the structurally related compounds demonstrate that the substitution of the aliphatic sidegroup at the R7 position has little effect on the overall ligand orientation (as shown for MAK15 and ligand 27 in Fig. 5.6).

When we now compute the changes in binding energy between associated pairs of ligands, we find that the R7 substitution with a aliphatic sidegroup improves the binding energy by 60kJ/mol on average.

For NAV478 the methyl to C9H19 substitution results in an improved binding energy by 89 kJ/mol. This supports an additive model of sidegroup contributions for the R7 group. Such an additive model for the contributions of the substituents is appropriate when (a) the substitution does not significantly affect the binding mode and (b) differences in the interactions are local. For the substitutions investigated here, the difference in binding energy arises mainly from changes in the solvent exposed surface of ligand and receptor, which results from changes in the entropy of the water displaced by the aliphatic sidegroup. On the basis of this model, one may postulate that all novel ligands should bind with a higher affinity than their structural analogues, with a similar order of the relative binding energy. In particular N478, which is derived from the compound with the best affinity from the set of Behrenswerth and has the strongest improvement of relative biding energy, should bind very strongly to CB1. Fig. 5.7 shows an excellent correlation of the observed affinities and the computed binding energies for this set of compounds, which demonstrates that this observation is in quantitative agreement with the experimental findings.

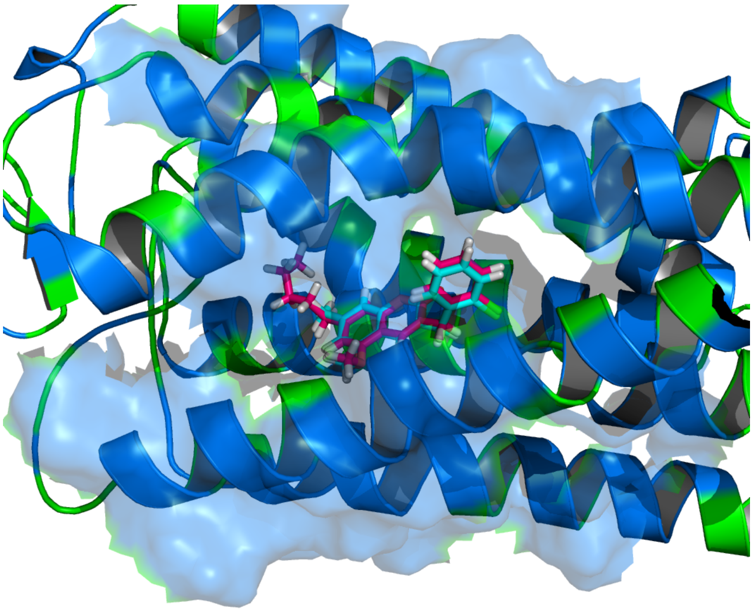


Figure 5.6: Binding poses between ligand 27 and MAK 15. The substitution of the aliphatic sidegroup at the R7 position has little effect on the overall binding pose.

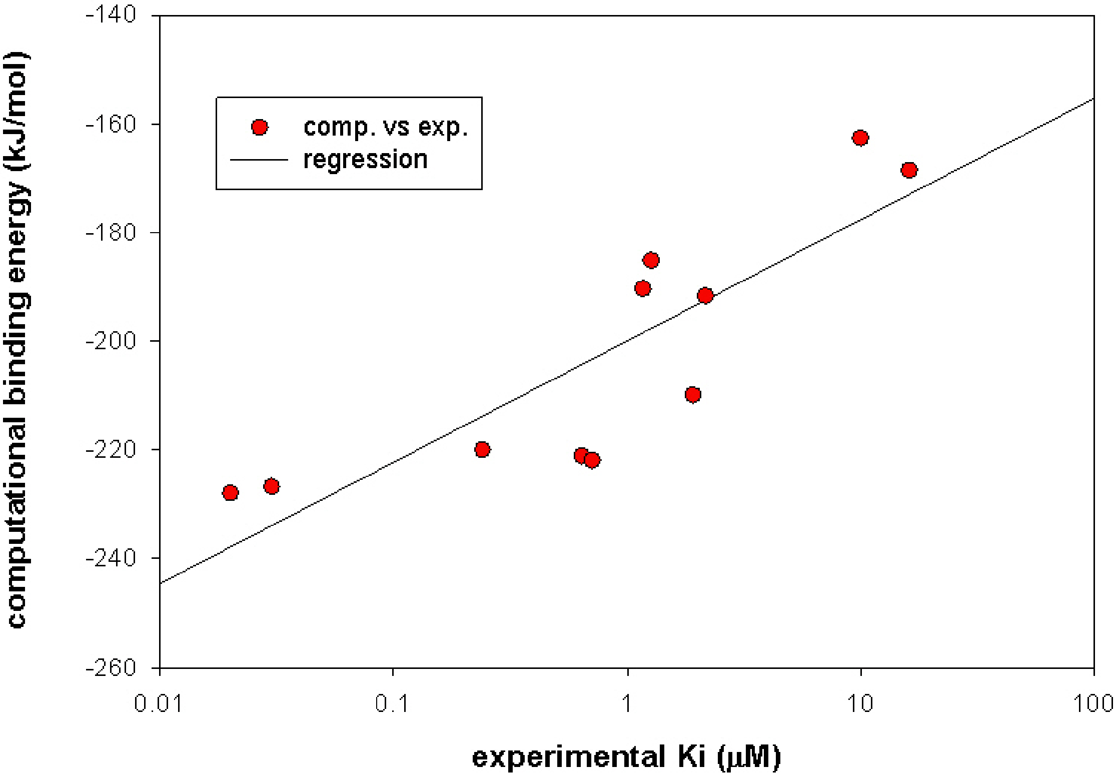


Figure 5.7: Plot correlating the results of the new ligands with the experimental one.

COMPOUND	BINDING ENERGY rigid with SASA	Rat CB1a versus [3H]CP55,940
CP55,940	-210,34	0.00124 ± 0.00058 (0.00137)
WIN55,212-2	-247,10	0.00606 ± 0.00062 (0.0044, 0.00994)
AM281	-176,20	0.0124 ± 0.0038 0.012
9	-165,32	>10 (18 ± 7%)
10	-171,49	4.19 ± 2.00
11	-195,23	>10 (41 ± 3%)
12	-161,61	>10 (10 ± 4%)
13	-166,36	19.1 ± 4.0
14	-161,02	>> 10 (8 ± 2%)
15	-119,76	>10 (15 ± 10%)
16	-216,48	>10 (31 ± 7%)
17	-143,01	>10 (12 ± 10%)
18	-108,53	>> 10 (8 ± 5%)
19	-120,68	>10 (32 ± 15%)
20	-162,89	>10 (31 ± 13%)
21	-132,36	>>10 (1 ± 6%)
22	-169,74	>>10 (6 ± 9%)
23	-130,79	>>10 (6 ± 3%)
24	-168,26	3.46 ± 1.14
25	-192,62	1.17 ± 0.32
26	-203,65	9.14 ± 1.87
27	-173,53	0.978 ± 0.264
28	-183,21	2.08 ± 0.40
29	-186,46	1.10 ± 0.12
30	-169,26	0.738 ± 0.414
31	-210,52	4.11 ± 1.38
32	-213,83	6.14 ± 0.69
33	-188,90	9.57 ± 2.00
34	-227,29	6.23 ± 1.43
35	-236,25	12.3 ± 3.2
36	-219,40	6.18 ± 0.96
37	-123,97	>10 (39 ± 6%)
38	-200,34	2.76 ± 0.66
39	-148,66	>10 (26 ± 11%)
40	-93,50	>10 (17 ± 6%)
41	-97,05	>>10 (46 ± 7%)
42	-152,66	>10 (34 ± 8%)
43	-170,33	4.90 ± 1.11
44	-184,69	1.37 ± 0.27

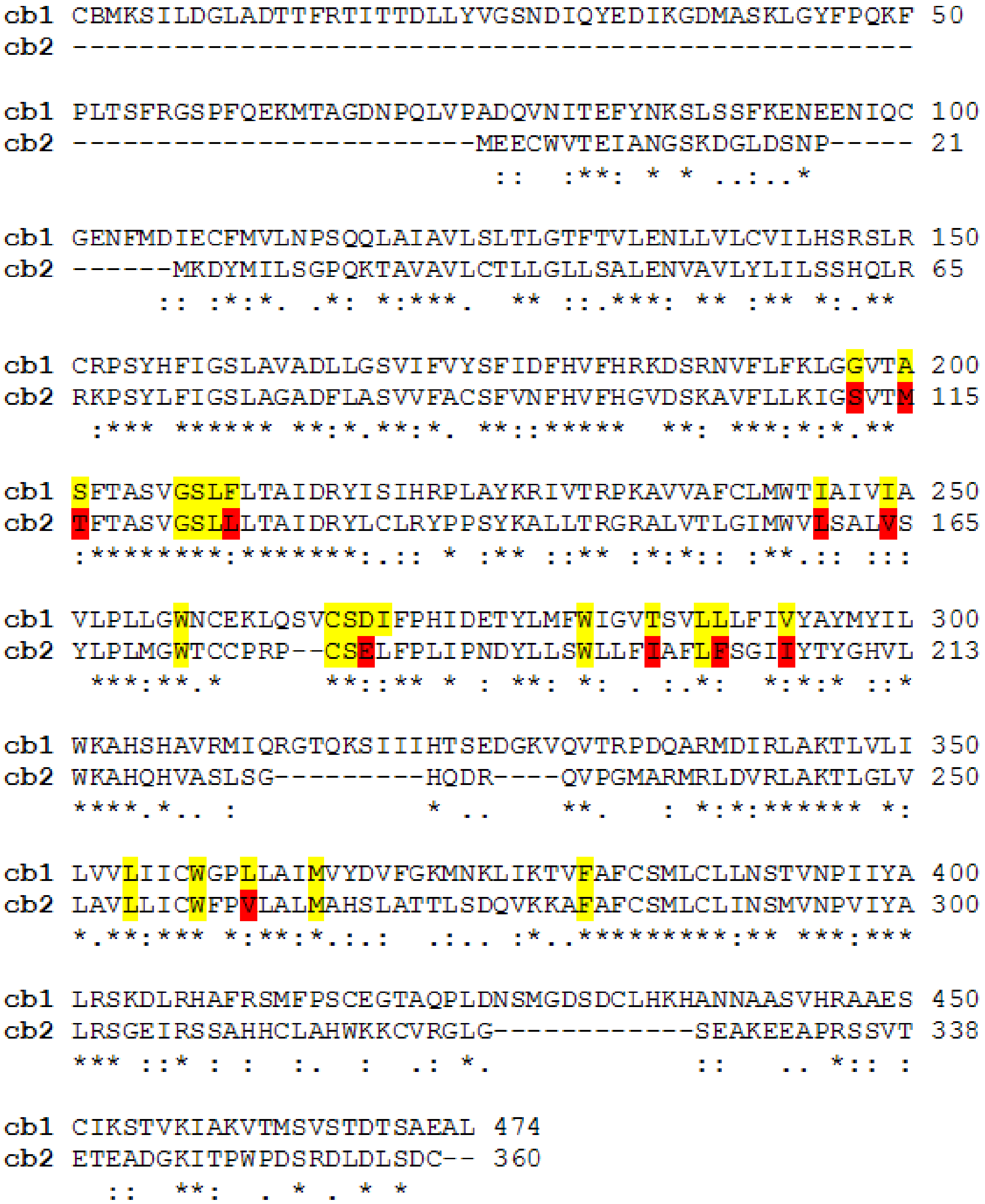
Table 5.II: Results for the docking simulations for CB1 compared with the experimental results (name of the compounds are in the APPENDIX).

5.4.3. Origins of CB1/CB2 selectivity

Receptor selectivity, which arises from minute differences in the interactions of different receptors with a given ligand, is difficult to understand even when high quality crystal structures are available. Many factors, including receptor conformational change, differences in long-range interactions and local factors contribute to receptor selectivity. Since our models for CB1/CB2 were built from alignments to the same template, the overall folds of the receptor models are very similar. We therefore investigated, whether selectivity could be explained on the basis of local effects alone and chose MAK15 and NV88 as the two most selective ligands in the test set as examples.

To investigate the importance of specific residues which differ between CB1 and CB2 and their contribution to receptor selectivity we have identified the set of residues (numbering 24 and illustrated in Fig 5.4A) of CB1 that are interacting with the ligands MAK15 and NV88 using Pymol (DeLano 2004). We then used ClustalW for sequence alignment of the CB1 to the CB2 (Fig 5.8) receptor to identify all 11 amino acids that differ in the interaction region between the two receptors, thereby narrowing the analysis to the local effects directly affecting the binding pocket (Fig 5.9).

For each of these 11 amino acids (Table 5.III) we prepared a model, where one specific amino acid in the CB1 structure was mutated to its corresponding amino acid in the CB2 receptor and one additional model (labeled model ALL), in which all 11 amino acids in the CB1 receptor were replaced by the corresponding amino acids in the CB2 receptor. The ALL model thus has replaces all interacting sidechains in the CB1 receptor with those found in CB2, while leaving the backbone conformation and all non-local differences between CB1 and CB2 unaltered. When we performed docking simulation of the selected ligands using the relaxation protocol described above with the ligands we find for NV88 that the single mutation in the CB1 receptor leave the binding energy (Table 5.IV) nearly unaffected. When we performed the relaxation in the ALL model we actually find an improvement of the binding energy, which is commensurate with the experimental finding (Table 5.IV). For the ligand MAK15 we find that most substitutions induce only small changes in the binding energy, but VAL293 and LEU361 induce a reduced affinity which is due to steric repulsion between the ligands and the increase of the size of the sidechain. Analyzing all mutations in the ALL model we find a significant reduction in the binding energy that agrees with the experimental result. These data suggest that for the class of ligands investigated here, receptor selectivity can be explained qualitatively by the specifics of the interactions of the different ligands with the amino acids in the direct vicinity of the binding pocket.



```
cb1 CBMKSILDGLADTTFRTITTDLLYVGSNDIQYEDIKGDMASKLGYFPQKF 50
cb2 --------------------------------------------------

cb1 PLTSFRGSPFQEKMTAGDNPQLVPADQVNITEFYNKSLSSFKENEENIQC 100
cb2 ------------------------MEECWVTEIANGSKDGLDSNP----- 21
                             ::  :**: * * ..:..*

cb1 GENFMDIECFMVLNPSQQLAIAVLSLTLGTFTVLENLLVLCVILHSRSLR 150
cb2 ------MKDYMILSGPQKTAVAVLCTLLGLLSALENVAVLYLILSSHQLR 65
          :: :*:*. .*: *:***.  ** ::.***: ** :** *:.**

cb1 CRPSYHFIGSLAVADLLGSVIFVYSFIDFHVFHRKDSRNVFLFKLGGVTA 200
cb2 RKPSYLFIGSLAGADFLASVVFACSFVNFHVFHGVDSKAVFLLKIGSVTM 115
     :*** ****** **:*.**:*. **::*****  **: ***:*:*.**

cb1 SFTASVGSLFLTAIDRYISIHRPLAYKRIVTRPKAVVAFCLMWTIAIVIA 250
cb2 TFTASVGSLLLTAIDRYLCLRYPPSYKALLTRGRALVTLGIMWVLSALVS 165
    :********:*******:.:: * :** ::** :*:*:: :**.:: :::

cb1 VLPLLGWNCEKLQSVCSDIFPHIDETYLMFWIGVTSVLLLFIVYAYMYIL 300
cb2 YLPLMGWTCCPRP--CSELFPLIPNDYLLSWLLFIAFLFSGIIYTYGHVL 213
     ***:**.*      **::** * : **: *: . :.*:  *:*:* ::*

cb1 WKAHSHAVRMIQRGTQKSIIIHTSEDGKVQVTRPDQARMDIRLAKTLVLI 350
cb2 WKAHQHVASLSG---------HQDR----QVPGMARMRLDVRLAKTLGLV 250
    ****.*.. :           * ..    **.   : *:*:****** *:

cb1 LVVLIICWGPLLAIMVYDVFGKMNKLIKTVFAFCSMLCLLNSTVNPIIYA 400
cb2 LAVLLICWFPVLALMAHSLATTLSDQVKKAFAFCSMLCLINSMVNPVIYA 300
    *.**:*** *:**:*.:.:  .:.. :*..*********:** ***:***

cb1 LRSKDLRHAFRSMFPSCEGTAQPLDNSMGDSDCLHKHANNAASVHRAAES 450
cb2 LRSGEIRSSAHHCLAHWKKCVRGLG------------SEAKEEAPRSSVT 338
    *** ::* : :  :.  :  .: *.            ::   .. *:: :

cb1 CIKSTVKIAKVTMSVSTDTSAEAL 474
cb2 ETEADGKITPWPDSRDLDLSDC-- 360
      ::  **:  . * . * *
```

Figure 5.8: Alignment between CB1 and CB2, in yellow we label the residues interacting with the ligands and in red those residues in the interaction region where CB1 and CB2 differ.

CB1 AA	CB1 mutated with CB2 AA
GLY (pos 197)	SER
ALA (pos 200)	MET
SER (pos 201)	THR
PHE (pos 210)	LEU
ILE (pos 245)	LEU
ILE (pos 249)	VAL
ASP (pos 268)	GLU
THR (pos 285)	ILE
LEU (pos 289)	PHE
VAL (pos 293)	ILE
LEU (pos 261)	VAL

Table 5.III: List of the 11 amino acids in the interaction region where CB1 differs from CB2.

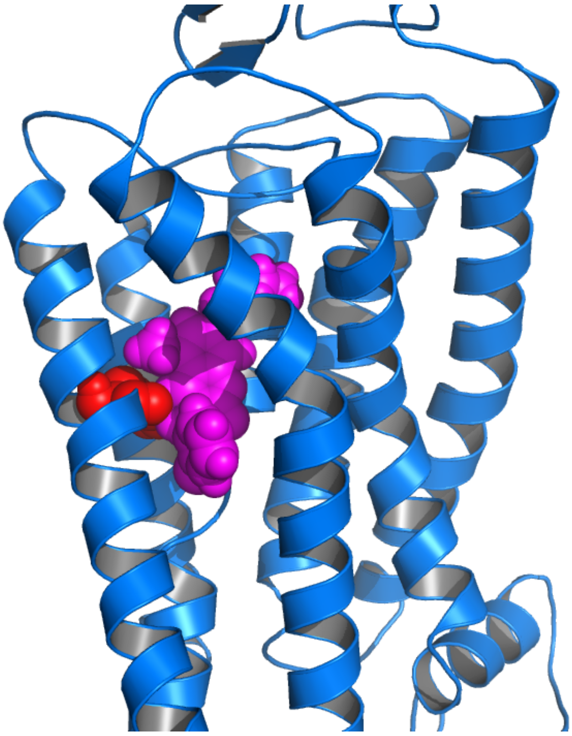


Figure 5.9: Local effects between MAK15 (magenta) and the mutated G197S (red) is directly affecting the binding pocket.

	BINDING ENERGY rigid with SASA CB1					
	NV88	BINDING ENERGY	ENERGY DIFFERENCE Kj/mol	MAK15	BINDING ENERGY	ENERGY DIFFERENCE Kj/mol
	WT	-234,71	0,00	**WT**	-211,47	0
MUTATED AA	**ALL MODELL**	-239,07	4,36	**ALL MODELL**	-208,59	-2,88
	G197S	-236,49	1,78	**G197S**	-214,45	2,98
	A200M	-235,24	0,53	**A200M**	-219,38	7,91
	S201T	-243,15	8,44	**S201T**	-220,35	8,88
	F210L	-234,98	0,27	**F210L**	-223,31	11,84
	I245L	-234,81	0,10	**I245L**	-223,31	11,84
	I249V	-234,47	-0,24	**I249V**	-221,27	9,8
	D268E	-234,09	-0,62	**D268E**	-223,31	11,84
	T285I	-234,97	0,26	**T285I**	-225,96	14,49
	L289F	-234,24	-0,47	**L289F**	-223,13	11,66
	V293I	-234,51	-0,20	**V293I**	-209,9	-1,57
	L361V	-234,92	0,21	**L361V**	-206,69	-4,78
	EXPERIMENTAL RESULTS					
	human CB_1 K_i (µM)	2,63 ± 1,23		**human CB_1 K_i (µM)**	0,033 ± 0,012	
	human CB_2 K_i (µM)	0,465 ± 0,024		**human CB_2 K_i (µM)**	0,185 ± 0,026	

Table 5.IV: Binding energy for the eleven single substitution models and the ALL model for the ligands MAK 15 and NV88.

5.4.4. Conclusions

In this study we wanted to contribute to an understanding of small-molecule binding to cannabinoid receptors, which have important roles in many diseases such as Huntington disorder and Alzheimer (Bisogno and Di Marzo 2010; Van Laere, Casteels et al. 2010). Rational design of novel inhibitors for these receptors is complicated by the fact that no crystal structures are available. I have therefore constructed models for the CB1 and CB2 receptors and then performed ligand-binding simulations for a family of 39 coumarine derivatives. We rationally designed, synthetized and tested a novel set of ligands to exploit a hydrophobic cavity in the vicinity of the docking pose of known ligands. All novel ligands showed improved, partially nanomolar, affinities with respect to their reference compounds, in good quantitative agreement between experiment and simulation, some of them reaching nanomolar affinities. Studying a series of site-specific mutations, we could computationally rationalize the receptor selectivity of specific compounds to either preferentially bind CB1 or CB2, respectively.

We used a special approach, by "walking" the CB1 receptor to its "CB2" counterpart by a series of single-point mutations of the interacting residues to understand the specificity for two important ligands.

5.5. Modulating activation of antithrombin in presence of heparin

In collaboration with Dr. Pérez-Sánchez and Prof. J. Corral at the University of Murcia we undertook a project trying to identify new molecules that may activate antithrombin using an *in silico* screening approach. Antithrombin is a member of the serpin superfamily of protease inhibitors. As other serpins, antithrombin inhibits its proteases by an unusual branched pathway suicide substrate mechanism in which the reactive centre loop of the inhibitor is cleaved by the protease as a normal substrate but is trapped as an acyl-intermediate covalent complex (Stratikos and Gettins 1999; Huntington, Read et al. 2000). However, antithrombin circulates in blood in a metastable conformation in which the reactive centre loop is partially inserted and is only activated by heparin and heparan sulfate glycosaminoglycans on the injured subendothelium (de Agostini, Watkins et al. 1990; Olson and Bjork 1994). Accordingly, sulfated polysaccharide heparin chains with different size, from unfractionated to the essential pentasaccharide, have been used successfully in anticoagulant therapy and thromboprophilaxis (Harenberg and Wehling 2008).

Heparin is probably one of the most widely used drugs in developed societies to prevent or treat thromboembolic diseases. Unfortunately, because of its highly anionic nature, heparin is involved in many interactions with blood proteins other than antithrombin and with vessel wall components (Conrad 1998), that can be involved in an appreciable individual variability in optimal dosage and an associated risk for bleeding complications. Moreover, such interactions may also associate with side effects like thrombocytopenia and osteoporosis (Devlin, Einhorn et al. 1988; Warkentin, Chong et al. 1998). These problems have inspired efforts to design safer and more specific therapeutic agents that would not be associated with such complications.

Since the discovery of the anticoagulant activity of heparins new molecules able to bind antithrombin have been identified. The strategies used in this search have been based mainly on the synthesis or chemical modification of existing drugs, or in the application of natural compounds with similar properties to the currently used compounds (Henry, Connell et al. 2009). That is the case of lignins and flavonoids (Gunnarsson and Desai 2004; Henry, Connell et al. 2009), highly sulfated small organic ligands which seem to have similar properties to heparins. An alternative approach is to screen a large database *in silico* and use affinity-ranking to identify some at least weakly-binding molecules for further refinement. Aided by ever-increasing computational power (Guerrero, Perez-Sanchez et al. 2011; Perez-Sanchez and Wenzel 2011), virtual screening is an appealing and cost-effective approach to tap into the wealth of available structural information (Ghosh, Nie et al. 2006). However, despite several success stories, limitations in current *in silico* screening approaches restrict their accuracy and general applicability (Klebe 2006; Warren, Andrews et al. 2006).

I contributed to this theoretical/experimental work, by performing an alanine scan of the Antithrombin/Heparin complex in order to provide an understanding of the interaction hot spots. With other members of our group I then screened a chemical library containing more than 13 millions of compounds *in silico* with the objective of seeking those capable of inducing a conformational change in antithrombin leading to its activation. Several ligands with high predictive affinity were subsequently validated experimentally, leading to a discovery of a compound with nanomolar affinity that acts as a cofactor to heparin to activate antithrombin *in vitro*, in mice and in human blood plasma (patent pending).

5.5.1. In-Silico Alanine Screen

Docking was performed against the crystal structure of antithrombin taken from the Antithrombin/Heparin pentasaccharide complex with PDB code 1AZX using FlexScreen protocol. We first validated the docking protocol by docking heparin to AT using the standard FlexScreen protocol. The simulation was divided into several partitions. In the first partition 500 simulations with 50000 computational steps, in the second partition 100 simulations with 30000 computational steps and in the third partition 10 simulations with 75000 computational steps were performed. The resulting structure of the complex (Fig. 5.11) showed very good similarity (RMSD 0.2 Å) with the native structure. Because heparin is a polysaccharide with 5 sugar rings, 149 atoms and 441 degrees of freedom it is a difficult ligand for *in silico* screening protocols.

This simulation was repeated by substituting all amino acid residues in the interaction region between antithrombin and heparin by alanine and the resulting binding energy was calculated to find out which residues of the receptor make the largest contribution to the binding energy (Fig. 5.10).

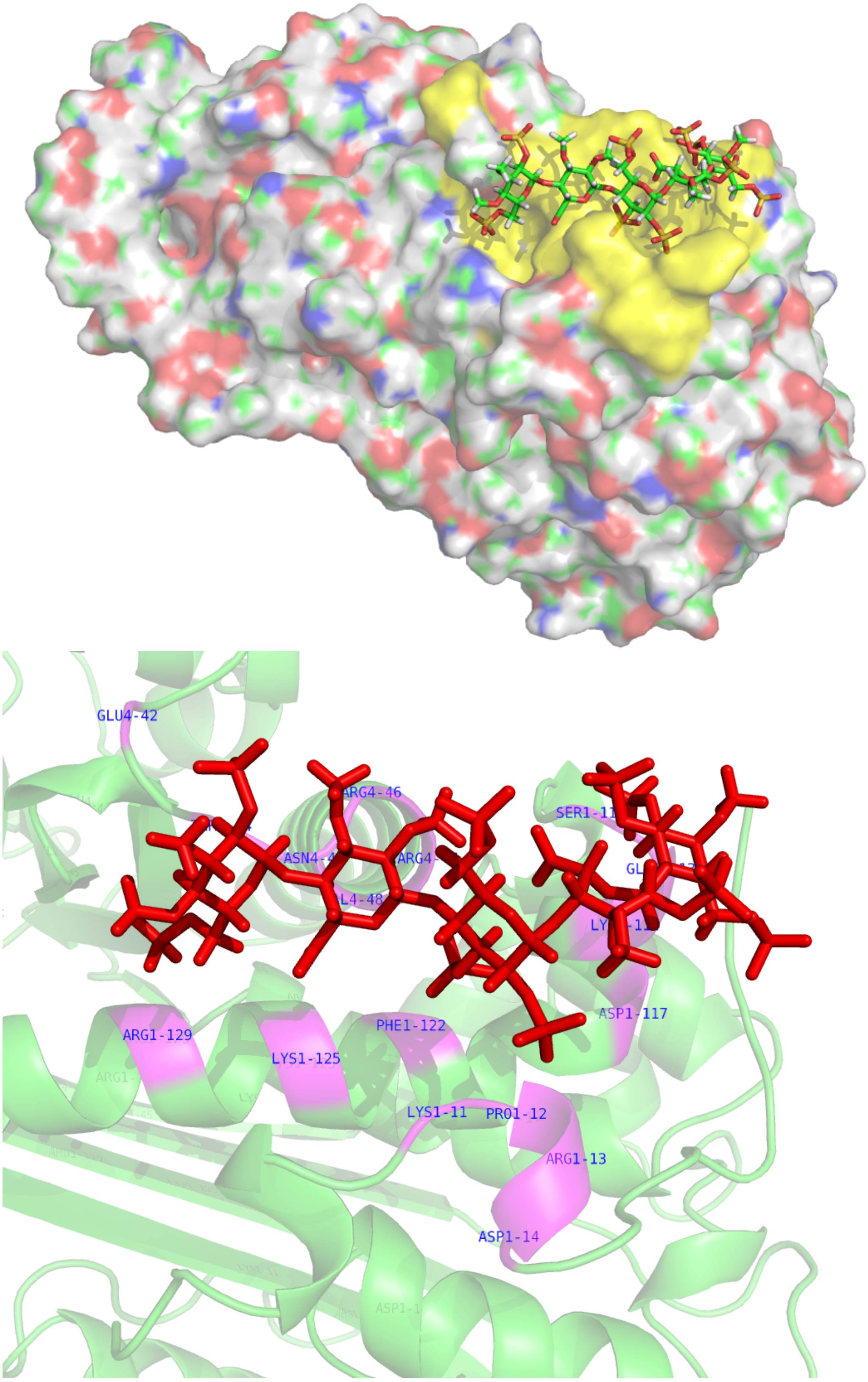


Figure 5.10: Top panel: annotated crystal structure of heparin (stick) and antithrombin (surface). Color-coded representation of antithrombin: oxygen (red), carbon (green), nitrogen (blue), hydrogen (white), and sulfur (yellow). The interface with heparin is hilighted in yellow in the antithrombin surface. Bottom panel: residues of antithrombin substituted with Alanine are highlighted in magenta (name of residues in blue).

Identification of such hotspots may help to define optimal docking pockets which may be target my small molecule ligands with a much lower molecular weight than heparin.

A total of 17 residues from Antithrombin were analyzed, among these eight mutations cause a significant difference of energy (Fig. 5.12 and Table 5.V).

These eight mutations are all charged residues, as LYS11, ARG46, ARG47, LYS114 and LYS125 which indicates strong electrostatic components for the stabilization of this complex. The substitution with non-polar amino acids with alanine show very small changes in energy compared to the native structure.

The dominant contribution of some residues for the interaction Antithrombin/Heparin guided our next experiments for developing new potential ligands.

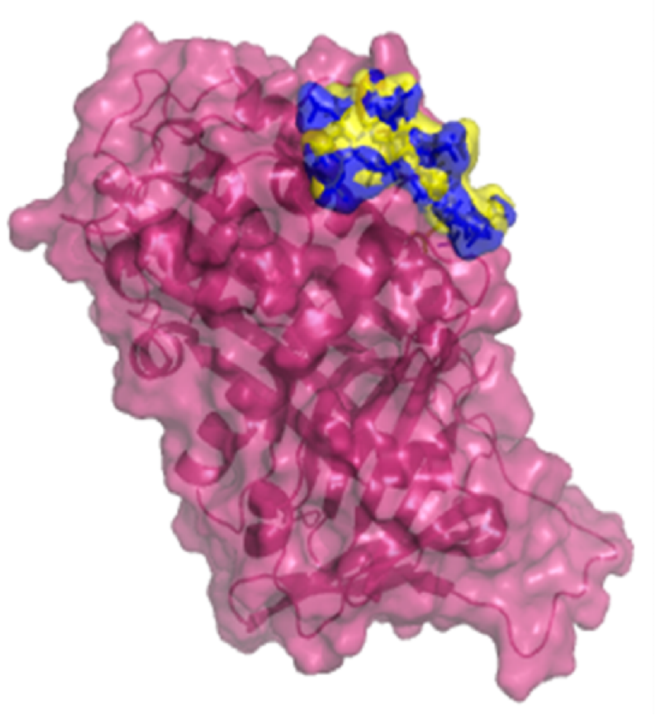


Figure 5.11: The predicted (yellow) and the experimental (blue) conformation of heparin in the complex between heparin and antithrombin (red).

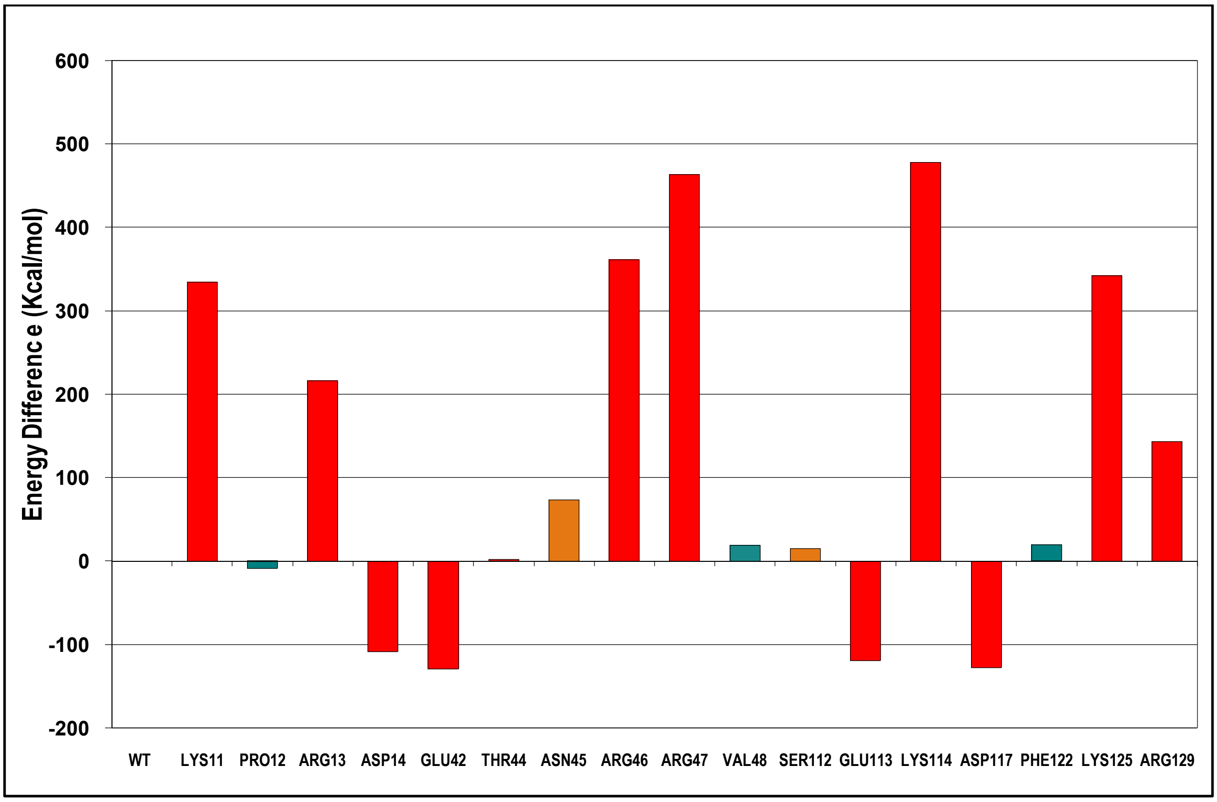


Figure 5.12: Histogram of the energy differences (in Kcal/mol) for all 17 residues of antithrombin complex with heparin. Colours of the columns identify different kinds of amino acids: red for AA with a polar acidic side chain; green for AA with a non polar side chain and orange for AA with a polar side chain.

Residue No.	Aminoacid	Ligand internal energy	Atom number	Number flex bonds	Binding Energy	External Energy	RMSD
	WT	1326,35	149,00	43,00	-2761,07	-2855,26	0,17
11	LYS	1326,35	149,00	43,00	-2426,27	-2530,61	0,13
12	PRO	1326,35	149,00	43,00	-2769,76	-2902,75	0,15
13	ARG	1326,35	149,00	43,00	-2544,32	-2649,12	0,17
14	ASP	1326,35	149,00	43,00	-2869,73	-3127,99	0,15
42	GLU	1326,35	149,00	43,00	-2889,99	-2973,20	0,14
44	THR	1326,35	149,00	43,00	-2758,51	-2945,04	0,12
45	ASN	1326,35	149,00	43,00	-2687,14	-2913,66	0,15
46	ARG	1326,35	149,00	43,00	-2399,51	-2520,01	0,14
47	ARG	1326,35	149,00	43,00	-2297,77	-2375,59	0,13
48	VAL	1326,35	149,00	43,00	-2742,01	-2969,62	0,15
112	SER	1326,35	149,00	43,00	-2745,89	-2897,87	0,15
113	GLU	1326,35	149,00	43,00	-2880,27	-2990,52	0,15
114	LYS	1326,35	149,00	43,00	-2282,62	-2392,42	0,19
117	ASP	1326,35	149,00	43,00	-2888,41	-3014,66	0,14
122	PHE	1326,35	149,00	43,00	-2741,70	-2841,97	0,14
125	LYS	1326,35	149,00	43,00	-2418,71	-2543,46	0,14
129	ARG	1326,35	149,00	43,00	-2617,40	-2800,15	0,16

Table 5.V: Energy contributions in (Kcal/mol) of alanine substitutions of individual amino acids in antithrombin.

5.5.2. In-Silico Screening

Motivated by these results, we developed an *in silico* screening protocol with the objective of seeking those capable of inducing a conformational change in antithrombin leading to its activation in a similar fashion to the one performed by heparin. We have therefore docked a large subset of chemical library containing more than 13 million of compounds.

We selected 8 ligands (Table 5.VI) with activating potential effect of antithrombin *in silico*. From the 8 selected ligands only six could be experimentally validated, since the other two were insoluble or were spontaneously hydrolyzed. From the six ligands tested, none showed a strong activation of anti-FXa and anti-FIIa activity of antithrombin. However, in presence of heparin, one of the tested compounds, D-myo-inositol 3,4,5,6-tetrakisphosphate (TMI), increased the heparin affinity of antithrombin in 2.3 fold, reducing its Kd to 45.35 nM. This increase in the affinity for heparin was correlated with higher anti-FXa and anti-FIIa activity of antithrombin in presence of this compound. Moreover, the selected compound together with heparin let to an activation of antithrombin in the plasma of patients with a type-II antithrombin deficiency in homozygosis (L99F) by 1.7 fold over the effect of heparin alone. This effect was not observed in the plasma of patients with type-II deficiency but with a different mutation (R47C) in the heparin binding site.

TMI could be used in activating antithrombin in patient as well as a part of a pharmaceutical composition for use in recovery of antithrombin activity in patients with antithrombin deficiency or in prophylaxis for a disease associated with increased blood clotting.

TMI activates antithrombin activity of thrombin inhibition by increasing the affinity of antithrombin to heparin. The affinity of heparin to antithrombin was measured using any standard fluorescence spectroscopic apparatus. Antithrombin contains four tryptophanes that are excited at 280 nm and produce the emission of fluorescence at 340 nm (W49, 189, 225, and 307). When heparin binds to antithrombin, antithrombin changes its conformation and is activated, which causes an increase in the intrinsic fluorescence of antithrombin. Since the fluorescence intensity of antithrombin increases proportionally to the amount of heparin bound, the affinity of antithrombin for the heparin can be calculated indirectly by measuring the emission of fluorescence in the presence of different concentrations of heparin until the affinity reaches saturation.

Compounds	Name of compound
Compound 1	4,5-dihidroxinaphatalene-2,7-disulfonic acid
Compound 2	1,2,4-benzeniltricarboxilic acid
Compound 3	Tetrapotassium 2,5-bis[(4-sulfonaftophenyl) diazenyl] benzene-1,3-disulfonate
Compound 4	D-myo-inositol 3,4,5,6-tetrakisphosphate
Compound 5	Disodium 4-hydroxy-5-sulfo-2,7-naphthalenedisulfonate
Compound 6	Cyclobutane-l,2,3,4-tetracarboxylic acid
Compound 7	4-[(4-carbpxyphenyl)carbamoyl]benzene-l,2-dicarboxylic acid
Compound 8	5-amino-4-hydroxy-8-sulphonaphthalene-2-sulfonat

Table 5.VI: 8 ligands selcted from *in silico* screening of 13 milion compounds that show an activating potential effect of antithrombin.

5.5.3. Conclusions

In this study we have pursued an *in silic*o discovery strategy in order to find non-polysaccharide scaffolding molecules from the ZINC-database with strong interactions with the heparin binding domain of antithrombin. The ligand with highest score D-myo-inositol 3,4,5,6-tetrakisphosphate (TMI), was experimentally validated confirming that this compound binds to antithrombin with nanomolar affinity, where TMI, interacts with even higher affinity than heparin. However, the identification of such interaction does not guarantee the same biochemical and functional consequences. Thus, TMI does not fully activate antithrombin as heparin does, and accordingly only slightly increases the rates of inhibition of thrombin and FXa. Interestingly, TMI increases the heparin affinity of antithrombin 2.3-15.2 fold. Finally, this molecule also expose the P1 residue of antithrombin (Arg393) allowing its citrullination at similar rates than heparin (Ordonez, Martinez-Martinez et al. 2009). All these data, suggest that TMI might induce a partially activated conformation (Johnson, Langdown et al. 2006). Actually, a partial activation of antithrombin should generate a conformational change that might facilitate the first step of the heparin binding (Olson, Schedin-Weiss et al. 2010) resulting in a higher affinity.

The increased activation of antithrombin together with the higher heparin affinity induced by TMI, and the absence of conformational side effects or negligible consequences on thrombin, sustain potential clinical usefulness for this molecule as a new anticoagulant drug. Indeed, this molecule might not significantly increase the risk of bleeding but it could allow higher or more efficient antithrombin activation by low affinity heparins, which are more abundant than high affinity heparins on vascular endothelium.

5.6. Summary

Protein-ligands interactions have an important role in processes in living systems. Studying of these interactions is of great interest, as it provides opportunities to understand protein function and therapeutic intervention. The accurate prediction of the binding modes between the ligand and protein, is of fundamental importance in modern structure-based drug design. With the studies reported in this chapter I presented examples of protein-ligand docking and a virtual screening of the ZINC-database using FlexScreen.

I have first performed a binding study of 19 small molecules synthetized by the group of Dr. Katja Schmitz with the chemokine receptor CXCR1. Results demonstrate that there are a few sites, where the molecules may weakly bind. One of the binding sites is near the position where interleukin-8 interacts with his receptor and could be a good candidate to be tested experimentally with the 4 small molecules found to interact strongest with this binding site.

In a second study I have constructed an homology model for CB1 and CB2 receptors and performed docking simulations for a set of compounds that have been investigated experimentally in the group of Prof. Stefan Bräse and Prof. Christa Müller. In this study I have constructed models for the CB1 and CB2 receptors demonstrating a successful approach of protein-ligand interaction using an homology modeling. I have performed ligand-binding simulations for a family of 39 coumarine derivatives, to finally design a novel set of ligands functionalized at position R7 of the coumarine scaffold with aliphatic sidegroups, which we also synthetized and tested experimentally. All novel ligands showed improved, partially nanomolar, affinities with respect to their reference compounds.

Finally in collaboration with the University of Murcia, we developed a structure-based model for AT-II regulation to develop new anticoagulation drugs. We used an *in silic*o strategy to screen the ZINC-database, to identify molecules able to bind to the heparin binding site of antithrombin. D-myo-inositol 3,4,5,6-tetrakisphosphate (TMI), which was identified in this screen, emerged as a promising candidate for experimental testing. The conformational change induced by TMI facilitated the interaction with heparin and more importantly with low affinity heparins. Thus, incubation of TMI with plasma of homozygous patients with antithrombin deficiency with heparin binding defect significantly improved the antithrombin inhibitory function. In conclusion, our *in silico* screening identified a new molecule able to interact with the heparin binding domain of antithrombin. The functional consequences of this interaction were experimentally characterized sustaining potential anticoagulant therapeutic applications.

6. Summary

Accelerating research in the life-sciences and adjacent fields, such as chemistry, physics and bioinformatics has led to a flood of data about biological systems, including humans, that for many perspectives help to better understand and influence them, and has led to the development of novel therapeutic strategies. One of the hallmarks of this development was the completion of the human genome project as the blueprint of our organism. Yet ten years after the completion of this project it has become evident that there are few diseases, or biological functions, that depend on the regulation of a single isolated gene. Indeed a picture has emerged in which proteins as the machines of the cell interact among each other and with specific other molecules, such as DNA, RNA and small-molecule ligands to form a network of regulatory circuits that control cell behavior.

In this context the study of biomolecular interactions has emerged as a key component of life-science research, as it connects the individual nodes of this network and permits their regulation. This raises the challenge to go beyond investigation of the individual components of cells and organisms in order to understand the network as a whole. Such knowledge is important to treat diseases and disorders that have no single cause and that involve multiple metabolic pathways. It is also important for novel life-science challenges, such as the design of new organisms, which is the goal of synthetic biology.

Fueled by both method development and an ever increasing availability of computational power, modeling techniques have increasingly complemented experiment in order to understand and modulate biological function. While the state-of-the-art of protein simulation studies in the mid-80s allowed modeling a single small protein with less than 5000 atoms for a few picoseconds it is now possible to model, with atomistic resolution, the structures of proteins and protein-complexes comprising hundreds of thousands of atoms for microseconds. Nevertheless these technologies are not yet mature enough to model protein behavior and protein-ligand interaction *de novo*, i.e. without recourse to experimental information. This has led to an increased and profitable cooperation between experiment and theory to speed-up biological discovery. The research field has also benefited tremendously from bioinformatics-based techniques that often operate at the sequence level, thereby exploiting directly the progress in gene-sequencing, to provide information about proteins and their interaction partners for which presently no structural information exists.

In this thesis I have reported studies that contributed in this exciting endeavor, modeling the interactions of proteins with other proteins and small-molecule ligands, in particular with applications in drug-discovery. It must be stressed that none of this work would have been possible without the support of experimental teams, and the vast set of bioinformatics-based resources that is now available in the internet. In addition to these resources I have used specific methods for protein-modeling and protein-ligand interaction that were developed in my group.

In the following I briefly summarize the results of the work reported:

Chemokine-receptor interactions

In this project I have studied the interaction of chemokines, regulatory proteins involved in inflammation, with their receptor in order to develop novel ligands that may influence chemokine-receptor interactions and thereby affect inflammatory disease (Meliciani, Klenin et al. 2009). For CXCL8/CXCR1 our method predicted seven important interaction hot spots in a small peptide derived from the N-terminus of CXCR1 that had been discovered experimentally (Baggiolini and Moser 1997; Skelton, Quan et al. 1999) and for the complex of the basolateral protein ERBB2 and the PDZ domain of its binding partner ERBIN, where we could recapitulate significant changes in interaction energy upon alanine substitution of a decisive tyrosine residue. Our simulation methodology yielded results that are in good agreement with experimental data showing that the prediction of hot spots from structural data is possible with a fraction of the computational effort needed for explicit molecular dynamics simulations of the protein-protein complex.

Cannabinoides

I have investigated the interaction between the Cannabinoid receptors 1/2 (CB1/CB2) that are members of the GCPR superfamily involved in neuroinflammatory and neurodegenerative disorders, with a novel set of coumarin derivatives. In cooperation with experimental partners we rationally designed, synthetized and tested a novel set of ligands to exploit a hydrophobic cavity in the vicinity of the docking pose of known ligands (Congress 2011). All novel ligands showed improved affinities with respect to their reference compounds in good quantitative agreement between experiment and simulation, some of them reaching nanomolar affinities. Studying a series of site-specific mutations, we could rationalize computationally the receptor selectivity of specific compounds to bind preferentially either CB1 or CB2. Together with our experimental colleagues we discovered new ligands with improved affinity to the CB1 receptor. Benchmarking our data with experiments we observed good agreement of the results and moreover this is a succesfull example of study using an homology modeling for protein/ligands interactions.

Antithrombin

Regulating blood coagulation is a very important challenge (other than a significant pharmaceutical market) relevant to many diseases, including coronary diseases, thrombosis and hemorrhage. In this project we aimed at developping new compounds able to activate antithrombin, a regulator of the blood coagulation cascade produced by the liver, using a new strategy structural docking. A chemical library containing more than 13 million compounds was screened with the objective of seeking those capable of inducing a conformational change in antithrombin leading to its activation in a similar fashion to the one performed by heparin. The results from this study revealed that structural docking could be a useful approach for searching new compounds with anticoagulant activity, since a new compound has been identified, TMI, with enhanced capacity for the activation of antithrombin in presence of its cofactor heparin. The TMI compound has potential application in the treatment of some antithrombin deficiencies (patent pending).

Existing compounds that can activate antithrombin are based on sugars, mainly heparin. Heparin may cause heparin-induced thrombocytopenia (HIT) in patients. Since heparin is only effective in the presence of antithrombin, the efficacy of heparin can be low if a patient has a low level of antithrombin activity. Other sugar-based antithrombin activators such as low molecular weight heparin (LMWH) and fondaparinux are mainly excreted in kidneys, and thus they can cause bleeding complications when administered to patients having kidney deficiency. The sugar-based antithrombin activators have also undesirable side effects, and in some cases they cannot be administrated to patients with sensitivity to sugars. In addition, sugars are usually very difficult to synthetize and therefore expensive, which limit their applicability as drugs. Other haemostatic elements have been also targets of common treatments for thrombosis and thromboembolism. Known drugs include antiplatelets (aspirin, clopidogrel, or anti-IIb/IIIa), anticoagulants (acenocoumarol, warfarin, anti-ll, or anti-FXa) and fibrinolytic drugs.

Application of structural docking as a tool for the discovery of new compounds with anticoagulant effects has the main advantage that it is possible to screen virtually million of compounds as a first filter, reducing considerably the experimental work. Our functional screening allowed us to identify a ligand acting as a promising coactivator of heparin, which has proven to be active *in vitro*, in mouse models and even in human blood plasma.

Protein structure prediction to identify causes for human developmental disease

One of the direct consequences of the success of the genome project is the availability of methods to rapidly sequence the genome of patients presenting a particular disease or disorder. Sometimes mutations are observed in a patient population afflicted with a particular disease. In order to translate these findings into therapeutic strategies is important to develop methods that can predict whether the observed mutations are causative for the observed phenotype. Despite all our efforts we are far from being able to make such predictions, as often the function of the proteins encoded by the afflicted genes is not known. Since protein structure determination is still very complex and costly, computational methods can be useful to head in such investigations.

This thesis reports work in a series of studies dealing with identification causes for Kallman syndrome/idiopathic hypogonadotropic hypogonadism, the identification of mutations in candidate genes assumes particular importance toward understanding the molecular basis of both nHH and KS. Given the genetic information on *CHD7*, *WDR11* and *NELF* we were able to investigate the possible functional effects of the mutations observed in the patients exhibiting the syndrome, using bioinformatics based methods and protein structure prediction to support the experimental work. *CHD7* represents the first identified chromatin-remodeling protein with a role in human puberty and the second gene to cause both normosmic IHH and KS in humans (Kim, Kurth et al. 2008). We were able to generate a model for the 3-D structure protein, for which no experimental structure was available. Our findings indicate that both normosmic IHH and KS are mild allelic variants of CHARGE syndrome and are caused by *CHD7* mutations. For *WDR11* I have also generated a model where our findings suggested that impaired pubertal development in patients results from a deficiency of productive *WDR11* protein interaction (Kim, Ahn et al. 2010). Finally for *NELF* our findings suggest that is associated with normosmic IHH and KS, either singly or in combination with a mutation in another gene (Xu, Kim et al. 2011).

With the work reported in this thesis I have attempted to contribute with computational techniques to the study of biomolecular interactions as one of the important regulators of setting a behavior. I hope that the results reported here demonstrate that bioinformatics and modeling techniques are mature enough to contribute to the elucidation of important biological phenomena. Presently none of these techniques work as a black-box, but the development of protocols requires careful choice, calibration and assessment of the available methods. The fraction of presently addressable problems constitutes only the tip of an iceberg of interesting biological questions awaiting further investigations. Much further method development and validation will be required to make more of these problems accessible. In the meantime some questions can be answered without close collaboration and input from experiment, requiring coordination among experimental and modeling groups to combine techniques from various fields in order to solve challenging problems. In addition the ever growing set of data available in the Internet emerges as an enormous resource to help researchers to address those questions.

With the growth of interdisciplinary collaborative projects and continuing investment in method development it is foreseeable that many questions that are difficult to approach today will become accessible in the future. The work presented here shows that by active collaborations experimentalists and theoreticians really do complement each other. As a matter of fact this thesis can be the starting point for further investigations focusing on modeling protein-protein and protein-ligand interactions. By also studying “hot spots” that can improve our understanding of fundamental biological processes and help to design lead structures in drug development, combining techniques such as docking, molecular dynamics, drug design, more detailed homology modeling and experimental work.

List of Figures

List of Tables

Compound Names referred to Table 5.I chapter 5.3 “Small Molecule Ligands for Binding Chemokines”, for the 19 small molecules

The IUPAC names are reported below:

COMPOUNDS	IUPAC name
Molecule 1	1-[dihydroxy(methyl)-λ^{8}-sulfanyl]-4- methylbenzene
Molecule 2	8-[(4-{[(4-chlorophenyl)dihydroxy-λ^{9}- sulfanylidene]amino}-1-oxidonaphthalen-2- yl)sulfanyl]-1λ^{1}-quinolin-1-uide
Molecule 3	2,6-bis({[(2E)-3-(furan-2-yl)-1-hydroxyprop-2-en-1- ylidene]amino})-1λ^{1}-pyridin-1-uide
Molecule 4	{[(5E)-1-hydroxyoctadec-5-en-7,9-diyn-1-yl]oxo}hydrogen
Molecule 5	3-(prop-2-en-1-yl)-1λ^{4},3-diazaspiro[4.5]decane-2,4-bis(ylium)-1-ide-2,4-bis(olate)
Molecule 6	(2R)-2-[(1E,3E)-penta-1,3-dien-1-yl]-2,3-dihydro-1-benzofuran-5,7-bis(olate)
Molecule 7	(2R,3S,4S,5R,6S)-2-{[3-(hydrogenylideneoxo)-1,3-dihydroxypropoxy]methyl}-6-{[(2S)-8-(3-methylbut-2-en-1-yl)-4,5-dioxido-2-(4-oxidophenyl)-3,4-dihydro-2H-1-benzopyran-7-yl]oxy}-1\$l^{3}-oxan-1-ide-3,4,5-tris(olate)
Molecule 8	3-(furan-3-ylmethylidene)-1,5-dioxaspiro[5.5]undecane-2,4,6-tris(ylium)-2,4-bis(olate)
Molecule 9	(2S,3R,4S,5S,6R)-2-{[(2S)-8-(3-methylbut-2-en-1- yl)-4,5-dioxido-2-(4-oxidophenyl)-3,4-dihydro-2H-1- benzopyran-7-yl]oxy}-6-(oxidomethyl)-1λ^{3}-oxan-1- ide-3,4,5-tris(olate)
Molecule 10	3-[(2E,6E)-8-(hydrogenylideneoxo)-8-hydroxy-4,4,7-trimethylocta-2,6-dien-1-yl]furan
Molecule 11	(2R,3R,4S,5S,6R)-2-{[(2S)-8-(3-methylbut-2-en-1-yl)-4,5-dioxido-2-phenyl-3,4-dihydro-2H-1-benzopyran-7-yl]oxy}-6-(oxidomethyl)oxan-1-uide-3,4,5-tris(olate)
Molecule 12	(2R,3S,4R,5R,6S)-2-O-[(3S,4R,5R,6S)-6-{[(2R,3R,4R,5R,6S)-6-{[(4aR,6aS,6bR,8aR,10S,12aR,12bR,14bS)-2,2,6a,6b,9,9,12a-heptamethyl-10-{[(2R,3R,4R,5S,6R)-6-methyl-3,4-dioxido-5-{[(2R,3S,4R,5S)-3,4,5-trioxido-1λ^{3}-oxan-1-id-2-yl]oxido}-1λ^{3}-oxan-1-id-2-yl]oxido}-1,2,3,4,4a,5,6,6a,6b,7,8,8a,9,10,11,12, 12a,12b,13,14b-icosahydropicen-4a-yl](hydroxy)methoxy}-4,5-dioxido-2-(oxidomethyl)-1\$l^{3}-oxan-1-id-3-yl]oxido}-5-oxido-4-{[(2R,3S,4R,5S)-3,4,5-trioxido-1\$l^{3}-oxan-1-id-2-yl]oxido}-1\$l^{3}-oxan-1-id-3-yl]-6-(oxidomethyl)-1\$l^{3}-oxan-1-ide-2,3,4,5-tetrakis(olate)
Molecule 13	(2S,3R,4S,5R,6R)-6-(oxidomethyl)-2-O-[(2R,3S)-4,5,7-trioxido-2-(3,4,5-trioxidophenyl)-3,4-dihydro-2H-1-benzopyran-3-yl]-1λ^{3}-oxan-1-ide- 2,3,4,5-tetrakis(olate)
Molecule 14	(2S)-6-methyl-2-(4-oxidophenyl)-3,4-dihydro-2H-1-benzopyran-4,5,7-tris(olate)
Molecule 15	{[(2,6-dimethylphenyl)imino](sulfanylidene)methylidene}(ethyl)amine
Molecule 16	3-(4-chlorophenyl)-5,5-dimethyl-4-methylidene-1,3-oxazolidin-2-ylium-2-olate
Molecule 17	(1S,3S,8S,9S,10S,13R)-8-[(1-hydroxy-2-methylprop-2- en-1-yl)oxy]-6,9,10-trimethyl-4,14-dioxatetracyclo [7.5.0.0^{1,13}.0^{3,7}]tetradec-6-en-3-ium-5-ylium-3,5-bis(olate)
Molecule 18	2-[(4bR,8S,8aS)-8-[(hydrogenylideneoxo)(hydroxy)methyl]-4b,8-dimethyl-4b,5,6,7,8,8a,9,10-octahydrophenanthren-2-yl]propan-2-olate
Molecule 19	(4S,8S)-4,8-dibenzyl-2,10-diphenyl- 1λ^{3},6,11λ^{3}-trioxa-3,9-diazaundeca-2,9-dien- 1,10-diyn-5-ol

Compound Names referred to Table 5.II chapter 5.4 “Selective high-affinity inhibitors of cannabinoid receptors CB1/CB2”, for the 3 reference compounds and the coumarine derivative. The IUPAC names are reported below:

COMPOUNDS	IUPAC name
CP55,940	(1S,3S,4S)-3-[2-hydroxy-4-2-methyloctan-2-yl)phenyl]-4-(3-oxidopropyl)cyclohexan-1-olate
WIN55,212-2	(11S)-2-methyl-11-(morpholin-4-ylmethyl)-3- [(naphthalen-1-yl)carbonyl]-9-oxa-1azatricyclo [6.3.1.0^{4,12}]dodeca-2,4(12),5,7- tetraene
AM281	1-(2,4-dichlorophenyl)-5-(4-iodophenyl)-4-methyl-N-(morpholin-4-yl)-1H-pyrazole-3-carboxamide
9	3-benzyl-2H-chromen-2-one
10	2-benzyl-3H-benzo[f]chromen-3-one
11	2-[(2-methoxyphenyl)methyl]-3H-benzo[f]chromen-3-one
12	2-methyl-3H-benzo[f]chromen-3-one
13	3-benzyl-5-methoxy-2H-chromen-2-one
14	3-benzyl-6-methoxy-2H-chromen-2-one
15	6-methoxy-3-methyl-2H-chromen-2-ylium-2-olate
16	3-benzyl-6-iodo-2H-chromen-2-one
17	6-iodo-3-methyl-2H-chromen-2-one
18	8-bromo-6-chloro-3-methyl-2H-chromen-2-one
19	3-benzyl-6-nitro-2H-chromen-2-one
20	3-benzyl-7-methoxy-2H-chromen-2-one
21	7-methoxy-3-methyl-2H-chromen-2-one
22	3-benzyl-8-methoxy-2H-chromen-2-ylium-2-olate
23	8-methoxy-3-methyl-2H-chromen-2-one
24	3-benzyl-5-methoxy-7-methyl-2H-chromen-2-one
25	5-methoxy-7-methyl-3-[(2-methylphenyl)methyl]-2H-chromen-2-one
26	5-methoxy-7-methyl-3-[(4-methylphenyl)methyl]-2H-chromen-2-one
27	3-[(2-chlorophenyl)methyl]-5-methoxy-7-methyl-2H-chromen-2-one
28	3-[(4-chlorophenyl)methyl]-5-methoxy-7-methyl-2H-chromen-2-one
29	3-[(2-hydroxyphenyl)methyl]-5-methoxy-7-methyl-2H-chromen-2-one
30	5-methoxy-3-[(2-methoxyphenyl)methyl]-7-methyl-2H-chromen-2-one
31	5-methoxy-3-[(3-methoxyphenyl)methyl]-7-methyl-2H-chromen-2-one
32	5-methoxy-3-[(4-methoxyphenyl)methyl]-7-methyl-2H-chromen-2-one
33	5-methoxy-3-[(2-methoxyphenyl)methyl]-2H-chromen-2-ylium-2-olate
34	3-[(2,4-dimethoxyphenyl)methyl]-5-methoxy-7-methyl-2H-chromen-2-one
35	5-methoxy-3-[(2-methoxyphenyl)methyl]-8-methyl-2H-chromen-2-one
36	3-[(2-methoxyphenyl)methyl]-8-methyl-5-(propan-2-yl)-2H-chromen-2-one
37	3-benzyl-8-bromo-5-methoxy-2H-chromen-2-one
38	3-benzyl-8-methyl-5-(propan-2-yl)-2H-chromen-2-one
39	3,8-dimethyl-5-(propan-2-yl)-2H-chromen-2-one
40	6-bromo-8-methoxy-3-methyl-2H-chromen-2-one
41	8-phosphanyl-2H-chromene-3-carbaldehyde
42	1-(8-bromo-5-methoxy-2-methyl-2H-chromen-3-yl)ethan-1-one
43	(2E,4aS)-2-[(2-oxidophenyl)methylidene]-2,3,4,4a-tetrahydro-1H-xanthen-1-ylium-3-id-1-olate
44	6-(1,3-dithiolan-2-yl)-2H-chromen-5-yl 1-ethyl(2Z)-2-[(2E)-3-phosphanylprop-2-en-1-ylidene]propanedioate

Adams, J. M. and S. Cory (1998). "The Bcl-2 protein family: arbiters of cell survival." Science **281**(5381): 1322-1326.

Agrafiotis, D. K., V. S. Lobanov, et al. (2002). "Combinatorial informatics in the post-genomics ERA." Nat Rev Drug Discov **1**(5): 337-346.

Aiyar, A. (2000). "The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment." Methods Mol Biol **132**: 221-241.

Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains." Science **181**(96): 223-230.

Ang, D., K. Liberek, et al. (1991). "Biological role and regulation of the universally conserved heat shock proteins." J Biol Chem **266**(36): 24233-24236.

Armougom, F., S. Moretti, et al. (2006). "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee." Nucleic Acids Res **34**(Web Server issue): W604-608.

Auld, D. S. (2001). "Zinc coordination sphere in biochemical zinc sites." Biometals **14**(3-4): 271-313.

Avbelj, F. and J. Moult (1995). "Role of electrostatic screening in determining protein main chain conformational preferences." Biochemistry **34**(3): 755-764.

Baggiolini, M. (1996). "Eotaxin: a VIC (very important chemokine) of allergic inflammation?" J Clin Invest **97**(3): 587.

Baggiolini, M. and B. Moser (1997). "Blocking chemokine receptors." J Exp Med **186**(8): 1189-1191.

Baker, D., S. Raman, et al. (2009). "Structure prediction for CASP8 with all-atom refinement using Rosetta." Proteins-Structure Function and Bioinformatics **77**: 89-99.

Barnhart, B. J. (1988). "The human genome project: a DOE perspective." Basic Life Sci **46**: 161-166.

Barth, F. and M. Rinaldi-Carmona (1999). "The development of cannabinoid antagonists." Curr Med Chem **6**(8): 745-755.

Bates, P. A., L. A. Kelley, et al. (2001). "Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM." Proteins **Suppl 5**: 39-46.

Battey, J. N. D., J. Kopp, et al. (2007). "Automated server predictions in CASP7." Proteins-Structure Function and Bioinformatics **69**: 68-82.

Behrenswerth, A., N. Volz, et al. (2009). "Synthesis and pharmacological evaluation of coumarin derivatives as cannabinoid receptor antagonists and inverse agonists." Bioorg Med Chem **17**(7): 2842-2851.

Benedix, A., C. M. Becker, et al. (2009). "Predicting free energy changes using structural ensembles." Nature Methods **6**(1): 3-4.

Benkert, P., S. C. E. Tosatto, et al. (2008). "QMEAN: A comprehensive scoring function for model quality assessment." Proteins-Structure Function and Bioinformatics **71**(1): 261-277.

Berman, H., K. Henrick, et al. (2007). "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data." Nucleic Acids Res **35**(Database issue): D301-303.

Bhagavath, B., R. H. Podolsky, et al. (2006). "Clinical and molecular characterization of a large sample of patients with hypogonadotropic hypogonadism." Fertil Steril **85**(3): 706-713.

Binder, K. (1996). "Introduction to Monte Carlo methods .1." Monte Carlo and Molecular Dynamics of Condensed Matter Systems **49**: 123-146.

Birrane, G., J. Chung, et al. (2003). "Novel mode of ligand recognition by the Erbin PDZ domain." J Biol Chem **278**(3): 1399-1402.

Bisogno, T. and V. Di Marzo (2010). "Cannabinoid receptors and endocannabinoids: role in neuroinflammatory and neurodegenerative disorders." CNS Neurol Disord Drug Targets **9**(5): 564-573.

Bissantz, C., B. Kuhn, et al. (2010). "A Medicinal Chemist's Guide to Molecular Interactions (vol 53, pg 5061, 2010)." Journal of Medicinal Chemistry **53**(16): 6241-6241.

Blake, K. D., N. Salem-Hartshorne, et al. (2005). "Adolescent and adult issues in CHARGE syndrome." Clin Pediatr (Phila) **44**(2): 151-159.

Blundell, T. L., B. L. Sibanda, et al. (1987). "Knowledge-based prediction of protein structures and the design of novel molecules." Nature **326**(6111): 347-352.

Bogan, A. A. and K. S. Thorn (1998). "Anatomy of hot spots in protein interfaces." Journal of Molecular Biology **280**(1): 1-9.

Borg, J. P., S. Marchetto, et al. (2000). "ERBIN: a basolateral PDZ protein that interacts with the mammalian ERBB2/HER2 receptor." Nat Cell Biol **2**(7): 407-414.

Boyd, S. (2005). "Molecular operating environment." Chemistry World **2**(9): 66-66.

Brooks, B. R., C. L. Brooks, et al. (2009). "CHARMM: The Biomolecular Simulation Program." Journal of Computational Chemistry **30**(10): 1545-1614.

Brown, T. A. (2002). Genomes. New York, Wiley-Liss.

Carroll, S. B., B. Prud'homme, et al. (2008). "Regulating evolution." Sci Am **298**(5): 60-67.

Cheatham, T. E., 3rd, J. Srinivasan, et al. (1998). "Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution." J Biomol Struct Dyn **16**(2): 265-280.

Chivian, D., D. E. Kim, et al. (2003). "Automated prediction of CASP-5 structures using the Robetta server." Proteins-Structure Function and Genetics **53**(6): 524-533.

Cho, K. I., D. Kim, et al. (2009). "A feature-based approach to modeling protein-protein interaction hot spots." Nucleic Acids Res **37**(8): 2672-2687.

Chothia, C. (1992). "Proteins. One thousand families for the molecular biologist." Nature **357**(6379): 543-544.

Chothia, C., A. M. Lesk, et al. (1986). "The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure." Science **233**(4765): 755-758.

Clackson, T. and J. A. Wells (1995). "A hot spot of binding energy in a hormone-receptor interface." Science **267**(5196): 383-386.

Collins, F. S. (1997). "Sequencing the human genome." Hosp Pract (Minneap) **32**(1): 35-43, 46-39, 53-34.

Congress (2011). "55th Biophysical Society meeting in Baltimore and at the 36th FEBS congress in Turin."

Conrad, H. E. (1998). Heparin-binding proteins. San Diego, Academic Press.

Cornell, W. D., P. Cieplak, et al. (1995). "A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules." Journal of the American Chemical Society **117**(19): 5179-5197.

Cui, M., M. Mezei, et al. (2008). "Prediction of protein loop structures using a local move Monte Carlo approach and a grid-based force field." Protein Engineering Design & Selection **21**(12): 729-735.

Dantas, G., C. Corrent, et al. (2007). "High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design." J Mol Biol **366**(4): 1209-1221.

Darnell, S. J., L. LeGault, et al. (2008). "KFC Server: interactive forecasting of protein interaction hot spots." Nucleic Acids Research **36**: W265-W269.

Darnell, S. J., D. Page, et al. (2007). "An automated decision-tree approach to predicting protein interaction hot spots." Proteins-Structure Function and Bioinformatics **68**(4): 813-823.

De Agostini, A. I., S. C. Watkins, et al. (1990). "Localization of anticoagulantly active heparan sulfate proteoglycans in vascular endothelium: antithrombin binding on cultured endothelial cells and perfused rat aorta." J Cell Biol **111**(3): 1293-1304.

De Genst, E., D. Areskoug, et al. (2002). "Kinetic and affinity predictions of a protein-protein interaction using multivariate experimental design." J Biol Chem **277**(33): 29897-29907.

De Roux, N., J. Young, et al. (1997). "A family with hypogonadotropic hypogonadism and mutations in the gonadotropin-releasing hormone receptor." N Engl J Med **337**(22): 1597-1602.

Delahaye, A., Y. Sznajer, et al. (2007). "Familial CHARGE syndrome because of CHD7 mutation: clinical intra- and interfamilial variability." Clin Genet **72**(2): 112-121.

DeLano, W. L. (2004). "Use of PYMOL as a communications tool for molecular science." Abstracts of Papers of the American Chemical Society **228**: U313-U314.

Devlin, V. J., T. A. Einhorn, et al. (1988). "Total hip arthroplasty after renal transplantation. Long-term follow-up study and assessment of metabolic bone status." J Arthroplasty **3**(3): 205-213.

Di Marzo, V. and I. Matias (2005). "Endocannabinoid control of food intake and energy balance." Nat Neurosci **8**(5): 585-589.

Dias, R., L. F. Timmers, et al. (2008). "Evaluation of molecular docking using polynomial empirical scoring functions." Curr Drug Targets **9**(12): 1062-1070.

Dillon, C., A. Creer, et al. (2002). "Basolateral targeting of ERBB2 is dependent on a novel bipartite juxtamembrane sorting signal but independent of the C-terminal ERBIN-binding domain." Mol Cell Biol **22**(18): 6553-6563.

DiMaio, F., T. C. Terwilliger, et al. (2011). "Improved molecular replacement by density- and energy-guided protein structure optimization." Nature **473**(7348): 540-543.

Dode, C., J. Levilliers, et al. (2003). "Loss-of-function mutations in FGFR1 cause autosomal dominant Kallmann syndrome." Nat Genet **33**(4): 463-465.

Dode, C., L. Teixeira, et al. (2006). "Kallmann syndrome: mutations in the genes encoding prokineticin-2 and prokineticin receptor-2." PLoS Genet **2**(10): e175.

Efremov, R., M. J. Truong, et al. (1999). "Human chemokine receptors CCR5, CCR3 and CCR2B share common polarity motif in the first extracellular loop with other human G-protein coupled receptors implications for HIV-1 coreceptor function." Eur J Biochem **263**(3): 746-756.

Eisenberg, D., W. Wilcox, et al. (1986). "Hydrophobicity and amphiphilicity in protein structure." J Cell Biochem **31**(1): 11-17.

Elofsson, A., T. Ohlson, et al. (2004). "Profile-profile methods provide improved fold-recognition: A study of different profile-profile alignment methods." Proteins-Structure Function and Bioinformatics **57**(1): 188-197.

Eramian, D., M. Y. Shen, et al. (2006). "A composite score for predicting errors in protein structure models." Protein Science **15**(7): 1653-1666.

Espadaler, J., N. Fernandez-Fuentes, et al. (2004). "ArchDB: automated protein loop classification as a tool for structural genomics." Nucleic Acids Res **32**(Database issue): D185-188.

Ewing, R. M., P. Chu, et al. (2007). "Large-scale mapping of human protein-protein interactions by mass spectrometry." Mol Syst Biol **3**: 89.

Fan, Y. X., L. Wong, et al. (2005). "EGFR kinase possesses a broad specificity for ErbB phosphorylation sites, and ligand increases catalytic-centre activity without affecting substrate binding affinity." Biochem J **392**(Pt 3): 417-423.

Felts, A. K., E. Gallicchio, et al. (2008). "Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling." Journal of Chemical Theory and Computation **4**(5): 855-868.

Fernandez-Fuentes, N., C. J. Madrid-Aliste, et al. (2007). "M4T: a comparative protein structure modeling server." Nucleic Acids Research **35**: W363-W368.

Fischer, B., S. Basili, et al. (2007). "Accuracy of binding mode prediction with a cascadic stochastic tunneling method." Proteins-Structure Function and Bioinformatics **68**(1): 195-204.

Fischer, B., H. Merlitz, et al. (2005). "Increasing diversity in in-silico screening with target flexibility." Computational Life Sciences, Proceedings **3695**: 186-197.

Fischer, T. B., K. V. Arunachalam, et al. (2003). "The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces." Bioinformatics **19**(11): 1453-1454.

Flanagan, J. F., B. J. Blus, et al. (2007). "Molecular implications of evolutionary differences in CHD double chromodomains." J Mol Biol **369**(2): 334-342.

Flanagan, S. E., A. M. Patch, et al. (2010). "Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations." Genet Test Mol Biomarkers **14**(4): 533-537.

Floudas, C. A., H. K. Fung, et al. (2006). "Advances in protein structure prediction and de novo protein design:Areview." Chemical Engineering Science **16**: 966 – 988.

Fong, H. K., J. B. Hurley, et al. (1986). "Repetitive segmental structure of the transducin beta subunit: homology with the CDC4 gene and identification of related mRNAs." Proc Natl Acad Sci U S A **83**(7): 2162-2166.

Forssmann, U., M. Uguccioni, et al. (1997). "Eotaxin-2, a novel CC chemokine that is selective for the chemokine receptor CCR3, and acts like eotaxin on human eosinophil and basophil leukocytes." J Exp Med **185**(12): 2171-2176.

Franke, W. W. (2004). "Actin's many actions start at the genes." Nat Cell Biol **6**(11): 1013-1014.

Friesner, R. A., J. L. Banks, et al. (2004). "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy." J Med Chem **47**(7): 1739-1749.

Frishman, D. and P. Argos (1995). "Knowledge-based protein secondary structure assignment." Proteins **23**(4): 566-579.

Fujitani, H., Y. Tanida, et al. (2005). "Direct calculation of the binding free energies of FKBP ligands." J Chem Phys **123**(8): 084108.

Gabb, H. A., R. M. Jackson, et al. (1997). "Modelling protein docking using shape complementarity, electrostatics and biochemical information." J Mol Biol **272**(1): 106-120.

Gao, X., D. Bu, et al. (2007). "FragQA: predicting local fragment quality of a sequence-structure alignment." Genome Inform **19**: 27-39.

Gerszten, R. E., E. A. Garcia-Zepeda, et al. (1999). "MCP-1 and IL-8 trigger firm adhesion of monocytes to vascular endothelium under flow conditions." Nature **398**(6729): 718-723.

Ghosh, S., A. Nie, et al. (2006). "Structure-based virtual screening of chemical libraries for drug discovery." Current Opinion in Chemical Biology **10**(3): 194-202.

Gilson, M. K. and H. X. Zhou (2007). "Calculation of protein-ligand binding affinities." Annu Rev Biophys Biomol Struct **36**: 21-42.

Ginalski, K. (2006). "Comparative modeling for protein structure prediction." Curr Opin Struct Biol **16**(2): 172-177.

Ginalski, K., A. Elofsson, et al. (2003). "3D-Jury: a simple approach to improve protein structure predictions." Bioinformatics **19**(8): 1015-1018.

Gohlke, H., C. Kiel, et al. (2003). "Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RaIGDS complexes." Journal of Molecular Biology **330**(4): 891-913.

Gohlke, H. and G. Klebe (2002). "Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors." Angew Chem Int Ed Engl **41**(15): 2644-2676.

Goodsell, D. S., R. Huey, et al. (2007). "A semiempirical free energy force field with charge-based desolvation." Journal of Computational Chemistry **28**(6): 1145-1152.

Goodsell, D. S., G. M. Morris, et al. (1996). "Automated docking of flexible ligands: applications of AutoDock." J Mol Recognit **9**(1): 1-5.

Gopal, S. M. and W. Wenzel (2006). "De novo folding of the DNA-binding ATF-2 zinc finger motif in an all-atom free-energy forcefield." Angew Chem Int Ed Engl **45**(46): 7726-7728.

Graves, D. T. and Y. Jiang (1995). "Chemokines, a family of chemotactic cytokines." Crit Rev Oral Biol Med **6**(2): 109-118.

Gross, C. (1996). "Function and Regulation of the Heat Shock Proteins, in Escherichia Coli and Salmonella: Cellular and Molecular Biology." ASM Press: 1384–1394.

Guerois, R., J. E. Nielsen, et al. (2002). "Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations." J Mol Biol **320**(2): 369-387.

Guerrero, G. D., H. Perez-Sanchez, et al. (2011). "Effective Parallelization of Non-bonded Interactions Kernel for Virtual Screening on GPUs." 5th International Conference on Practical Applications of Computational Biology & Bioinformatics (Pacbb 2011) **93**: 63-69.

Guimaraes, C. R., D. L. Boger, et al. (2005). "Elucidation of fatty acid amide hydrolase inhibition by potent alpha-ketoheterocycle derivatives from Monte Carlo simulations." J Am Chem Soc **127**(49): 17377-17384.

Gunnarsson, G. T. and U. R. Desai (2004). "Hydropathic interaction analyses of small organic activators binding to antithrombin." Bioorg Med Chem **12**(3): 633-640.

Halperin, I., H. Wolfson, et al. (2004). "Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking." Structure **12**(6): 1027-1038.

Hansson, T., J. Marelius, et al. (1998). "Ligand binding affinity prediction by linear interaction energy methods." J Comput Aided Mol Des **12**(1): 27-35.

Harenberg, J. and M. Wehling (2008). "Current and future prospects for anticoagulant therapy: inhibitors of factor Xa and factor IIa." Semin Thromb Hemost **34**(1): 39-57.

Hartmann, C., I. Antes, et al. (2007). "IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models." Protein Science **16**(7): 1294-1307.

He, M. M., A. S. Smith, et al. (2005). "Small-molecule inhibition of TNF-alpha." Science **310**(5750): 1022-1025.

Heath, H., S. Qin, et al. (1997). "Chemokine receptor usage by human eosinophils. The importance of CCR3 demonstrated using an antagonistic monoclonal antibody." J Clin Invest **99**(2): 178-184.

Henry, B. L., J. Connell, et al. (2009). "Interaction of antithrombin with sulfated, low molecular weight lignins: opportunities for potent, selective modulation of antithrombin function." J Biol Chem **284**(31): 20897-20908.

Herges, T. and W. Wenzel (2004). "An all-atom force field for tertiary structure prediction of helical proteins." Biophys J **87**(5): 3100-3109.

Higa, L. A. and H. Zhang (2007). "Stealing the spotlight: CUL4-DDB1 ubiquitin ligase docks WD40-repeat proteins to destroy." Cell Div **2**: 5.

Higgins, D. G. and P. M. Sharp (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer." Gene **73**(1): 237-244.

Higgins, D. G., J. D. Thompson, et al. (1996). "Using CLUSTAL for multiple sequence alignments." Methods Enzymol **266**: 383-402.

Hill, S. J. (2006). "G-protein-coupled receptors: past, present and future." Br J Pharmacol **147 Suppl 1**: S27-37.

Hogeweg, P. (1978). "Simulating Growth of Cellular Forms." Simulation **31**(3): 90-96.

Honig, B., C. L. Tang, et al. (2003). "On the role of structural information in remote homology detection and sequence alignment: New methods using hybrid sequence profiles." Journal of Molecular Biology **334**(5): 1043-1062.

Horuk, R., A. W. Martin, et al. (1997). "Expression of chemokine receptors by subsets of neurons in the central nervous system." J Immunol **158**(6): 2882-2890.

Hu, Z. J., B. Y. Ma, et al. (2000). "Conservation of polar residues as hot spots at protein interfaces." Proteins-Structure Function and Genetics **39**(4): 331-342.

Huang, G. T., S. K. Haake, et al. (1998). "Gingival epithelial cells increase interleukin-8 secretion in response to Actinobacillus actinomycetemcomitans challenge." J Periodontol **69**(10): 1105-1110.

Huang, N., B. K. Shoichet, et al. (2006). "Benchmark sets for molecular docking." J. Med. Chem. **49**: 6789-6801.

Humbles, A. A., B. Lu, et al. (2002). "The murine CCR3 receptor regulates both the role of eosinophils and mast cells in allergen-induced airway inflammation and hyperresponsiveness." Proc Natl Acad Sci U S A **99**(3): 1479-1484.

Huntington, J. A., R. J. Read, et al. (2000). "Structure of a serpin-protease complex shows inhibition by deformation." Nature **407**(6806): 923-926.

Huo, S., I. Massova, et al. (2002). "Computational alanine scanning of the 1 : 1 human growth hormone-receptor complex." Journal of Computational Chemistry **23**(1): 15-27.

Ito, T., T. Chiba, et al. (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc Natl Acad Sci U S A **98**(8): 4569-4574.

Jacobson, M. P., D. L. Pincus, et al. (2004). "A hierarchical approach to all-atom protein loop prediction." Proteins **55**(2): 351-367.

Jain, E., A. Bairoch, et al. (2009). "Infrastructure for the life sciences: design and implementation of the UniProt website." BMC Bioinformatics **10**: -.

Jaulin-Bastard, F., H. Saito, et al. (2001). "The ERBB2/HER2 receptor differentially interacts with ERBIN and PICK1 PSD-95/DLG/ZO-1 domain proteins." J Biol Chem **276**(18): 15256-15263.

Jeong, H., B. Tombor, et al. (2000). "The large-scale organization of metabolic networks." Nature **407**(6804): 651-654.

Johnson, D. J., J. Langdown, et al. (2006). "Crystal structure of monomeric native antithrombin reveals a novel reactive center loop conformation." J Biol Chem **281**(46): 35478-35486.

Jones, D. O., I. G. Cowell, et al. (2000). "Mammalian chromodomain proteins: their role in genome organisation and expression." Bioessays **22**(2): 124-137.

Jones, G., P. Willett, et al. (1997). "Development and validation of a genetic algorithm for flexible docking." J Mol Biol **267**(3): 727-748.

Jongmans, M. C., L. H. Hoefsloot, et al. (2008). "Familial CHARGE syndrome and the CHD7 gene: a recurrent missense mutation, intrafamilial recurrence and variability." Am J Med Genet A **146A**(1): 43-50.

Jorgensen, W. L. (1991). "Rusting of the lock and key model for protein-ligand binding." Science **254**(5034): 954-955.

Jorgensen, W. L. and M. N.A. (1997). "Development of an all-atomic force field for heterocycles properties of liquid pyridine and diazenes." J. Mol. Struct. **424**: 145.

Jose, P. J., I. M. Adcock, et al. (1994). "Eotaxin: cloning of an eosinophil chemoattractant cytokine and increased mRNA expression in allergen-challenged guinea-pig lungs." Biochem Biophys Res Commun **205**(1): 788-794.

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

Kelley, L. A. and M. J. Sternberg (2009). "Protein structure prediction on the Web: a case study using the Phyre server." Nat Protoc **4**(3): 363-371.

Keskin, O., B. Y. Ma, et al. (2005). "Hot regions in protein-protein interactions: The organization and contribution of structurally conserved hot spot residues." Journal of Molecular Biology **345**(5): 1281-1294.

Kim, H. G., J. W. Ahn, et al. (2010). "WDR11, a WD protein that interacts with transcription factor EMX1, is mutated in idiopathic hypogonadotropic hypogonadism and Kallmann syndrome." Am J Hum Genet **87**(4): 465-479.

Kim, H. G., B. Bhagavath, et al. (2008). "Clinical Manifestations of Impaired GnRH Neuron Development and Function." Neurosignals **16**(2-3): 165-182.

Kim, H. G., S. R. Herrick, et al. (2005). "Hypogonadotropic hypogonadism and cleft lip and palate caused by a balanced translocation producing haploinsufficiency for FGFR1." J Med Genet **42**(8): 666-672.

Kim, H. G., I. Kurth, et al. (2008). "Mutations in CHD7, encoding a chromatin-remodeling protein, cause idiopathic hypogonadotropic hypogonadism and Kallmann syndrome." Am J Hum Genet **83**(4): 511-519.

Kitano, H. (2002). "Computational systems biology." Nature **420**(6912): 206-210.

Kitano, H. (2002). "Systems biology: a brief overview." Science **295**(5560): 1662-1664.

Kitano, H. (2003). "[Introductions to systems biology]." Tanpakushitsu Kakusan Koso **48**(7): 789-793.

Kitaura, M., N. Suzuki, et al. (1999). "Molecular cloning of a novel human CC chemokine (Eotaxin-3) that is a functional ligand of CC chemokine receptor 3." J Biol Chem **274**(39): 27975-27980.

Kitchen, D. B., H. Decornez, et al. (2004). "Docking and scoring in virtual screening for drug discovery: methods and applications." Nat Rev Drug Discov **3**(11): 935-949.

Klabunde, T. (2007). "Chemogenomic approaches to drug discovery: similar receptors bind similar ligands." Br J Pharmacol **152**(1): 5-7.

Klebe, G. (2006). "Virtual ligand screening: strategies, perspectives and limitations." Drug Discov Today **11**(13-14): 580-594.

Kokh, D. B. and W. Wenzel (2008). "Flexible side chain models improve enrichment rates in in silico screening." J Med Chem **51**(19): 5919-5931.

Kortemme, T. and D. Baker (2002). "A simple physical model for binding energy hot spots in protein-protein complexes." Proc Natl Acad Sci U S A **99**(22): 14116-14121.

Kramer, P. R. and S. Wray (2000). "Novel gene expressed in nasal region influences outgrowth of olfactory axons and migration of luteinizing hormone-releasing hormone (LHRH) neurons." Genes Dev **14**(14): 1824-1834.

Kramer, P. R. and S. Wray (2001). "Nasal embryonic LHRH factor (NELF) expression within the CNS and PNS of the rodent." Brain Res Gene Expr Patterns **1**(1): 23-26.

Kryshtafovych, A., K. Fidelis, et al. (2007). "Progress from CASP6 to CASP7." Proteins-Structure Function and Bioinformatics **69**: 194-207.

Kryshtafovych, A., K. Fidelis, et al. (2009). "CASP8 results in context of previous experiments." Proteins-Structure Function and Bioinformatics **77**: 217-228.

Kudryashov, D. S., M. R. Sawaya, et al. (2005). "The crystal structure of a cross-linked actin dimer suggests a detailed molecular interface in F-actin." Proc Natl Acad Sci U S A **102**(37): 13105-13110.

Kulharia, M., R. S. Goody, et al. (2008). "Information theory-based scoring function for the structure-based prediction of protein-ligand binding affinity." Journal of Chemical Information and Modeling **48**(10): 1990-1998.

Larsson, P., B. Wallner, et al. (2008). "Using multiple templates to improve quality of homology models in automated homology modeling." Protein Science **17**(6): 990-1002.

Laurie, A. T. and R. M. Jackson (2005). "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites." Bioinformatics **21**(9): 1908-1916.

Lawrence, M. B. and T. A. Springer (1991). "Leukocytes roll on a selectin at physiologic flow rates: distinction from and prerequisite for adhesion through integrins." Cell **65**(5): 859-873.

Layman, L. C., D. P. Cohen, et al. (1998). "Mutations in gonadotropin-releasing hormone receptor gene cause hypogonadotropic hypogonadism." Nat Genet **18**(1): 14-15.

Leach, A. R. (2001). Molecular modelling : principles and applications. Harlow, England; New York, Prentice Hall.

Lee, B. and F. M. Richards (1971). "The interpretation of protein structures: estimation of static accessibility." J Mol Biol **55**(3): 379-400.

Lee, D. S., C. Seok, et al. (2008). "Protein loop modeling using fragment assembly." Journal of the Korean Physical Society **52**(4): 1137-1142.

Lee, F. S., Z. T. Chu, et al. (1992). "Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603." Protein Engineering **5**(3): 215-228.

Lengauer, T. and M. Rarey (1996). "Computational methods for biomolecular docking." Curr Opin Struct Biol **6**(3): 402-406.

Lensink, M. F., R. Mendez, et al. (2007). "Docking and scoring protein complexes: CAPRI 3rd Edition." Proteins **69**(4): 704-718.

Levitt, M. (1992). "Accurate modeling of protein conformation by automatic segment matching." J Mol Biol **226**(2): 507-533.

Li, D. and R. Roberts (2001). "WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases." Cell Mol Life Sci **58**(14): 2085-2097.

Linnarsson, S. (2010). "Recent advances in DNA sequencing methods - general principles of sample preparation." Exp Cell Res **316**(8): 1339-1343.

Liu, H., Y. Hwangbo, et al. (2004). "Analysis of genetic polymorphisms in CCR5, CCR2, stromal cell-derived factor-1, RANTES, and dendritic cell-specific intercellular adhesion molecule-3-grabbing nonintegrin in seronegative individuals repeatedly exposed to HIV-1." J Infect Dis **190**(6): 1055-1058.

Lu, M. Y., A. D. Dousis, et al. (2008). "OPUS-Rota: A fast and accurate method for side-chain modeling." Protein Science **17**(9): 1576-1585.

Luria, S. E. (1989). "Human genome program." Science **246**(4932): 873-874.

Ma, B. Y., T. Elkayam, et al. (2003). "Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces." Proceedings of the National Academy of Sciences of the United States of America **100**(10): 5772-5777.

Ma, W., P. J. Bryce, et al. (2002). "CCR3 is essential for skin eosinophilia and airway hyperresponsiveness in a murine model of allergic skin inflammation." J Clin Invest **109**(5): 621-628.

Mandell, J. G., V. A. Roberts, et al. (2001). "Protein docking using continuum electrostatics and geometric fit." Protein Engineering **14**(2): 105-113.

Marrero-Ponce, Y., R. Medina-Marrero, et al. (2005). "Protein linear indices of the 'macromolecular pseudograph alpha-carbon atom adjacency matrix' in bioinformatics. Part 1: Prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor." Bioorganic & Medicinal Chemistry **13**(8): 3003-3015.

Marti-Renom, M. A., M. S. Madhusudhan, et al. (2004). "Alignment of protein sequences by their profiles." Protein Science **13**(4): 1071-1087.

Massova, I. and P. A. Kollman (1999). "Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies." Journal of the American Chemical Society **121**(36): 8133-8143.

Matsuda, L. A., S. J. Lolait, et al. (1990). "Structure of a cannabinoid receptor and functional expression of the cloned cDNA." Nature **346**(6284): 561-564.

Matthews, B. W., B. Q. Wei, et al. (2004). "Testing a flexible-receptor docking algorithm in a model binding site." Journal of Molecular Biology **337**(5): 1161-1182.

McGuffin, L. J. (2007). "Benchmarking consensus model quality assessment for protein fold recognition." BMC Bioinformatics **8**: -.

McGuffin, L. J. (2008). "The ModFOLD server for the quality assessment of protein structural models." Bioinformatics **24**(4): 586-587.

Meliciani, I., K. Klenin, et al. (2009). "Probing hot spots on protein-protein interfaces with all-atom free-energy simulation." J Chem Phys **131**(3): 034114.

Mereghetti, P., M. L. Ganadu, et al. (2008). "Validation of protein models by a neural network approach." BMC Bioinformatics **9**: -.

Merlitz, H., B. Burghardt, et al. (2003). "Stochastic tunneling method for high throughput database screening." Nanotech 2003, Vol 1: 44-47

Merlitz, H. and W. Wenzel (2004). "High throughput in-silico screening against flexible protein receptors." Computational Science and Its Applications - Iccsa 2004, Pt 3 **3045**: 465-472.

Metzner, B., C. Hofmann, et al. (1999). "Overexpression of CXC-chemokines and CXC-chemokine receptor type II constitute an autocrine growth mechanism in the epidermoid carcinoma cells KB and A431." Oncol Rep **6**(6): 1405-1410.

Michel, J., M. L. Verdonk, et al. (2006). "Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization?" J Med Chem **49**(25): 7427-7439.

Miller, L. J., S. H. Kurtzman, et al. (1998). "Expression of interleukin-8 receptors on tumor cells and vascular endothelial cells in human breast cancer tissue." Anticancer Res **18**(1A): 77-81.

Miura, K., J. S. Acierno, Jr., et al. (2004). "Characterization of the human nasal embryonic LHRH factor gene, NELF, and a mutation screening among 65 patients with idiopathic hypogonadotropic hypogonadism (IHH)." J Hum Genet **49**(5): 265-268.

Mobley, D. L. and K. A. Dill (2009). "Binding of small-molecule ligands to proteins: "what you see" is not always "what you get"." Structure **17**(4): 489-498.

Mohri, K., S. Vorobiev, et al. (2004). "Identification of functional residues on Caenorhabditis elegans actin-interacting protein 1 (UNC-78) for disassembly of actin depolymerizing factor/cofilin-bound actin filaments." J Biol Chem **279**(30): 31697-31707.

Montgomerie, S., J. A. Cruz, et al. (2008). "PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation." Nucleic Acids Res **36**(Web Server issue): W202-209.

Mulder, N. J. and R. Apweiler (2008). "The InterPro database and tools for protein domain analysis." Curr Protoc Bioinformatics **Chapter 2**: Unit 2 7.

Mulder, N. J., R. Apweiler, et al. (2003). "The InterPro Database, 2003 brings increased coverage and new features." Nucleic Acids Res **31**(1): 315-318.

Murdoch, C., P. N. Monk, et al. (1999). "Cxc chemokine receptor expression on human endothelial cells." Cytokine **11**(9): 704-712.

Ng, P. C. and S. Henikoff (2003). "SIFT: Predicting amino acid changes that affect protein function." Nucleic Acids Res **31**(13): 3812-3814.

Nickel, R., L. A. Beck, et al. (1999). "Chemokines and allergic disease." J Allergy Clin Immunol **104**(4 Pt 1): 723-742.

Norgauer, J., B. Metzner, et al. (1996). "Expression and growth-promoting function of the IL-8 receptor beta in human melanoma cells." J Immunol **156**(3): 1132-1137.

Obiol-Pardo, C. and J. Rubio-Martinez (2007). "Comparative evaluation of MMPBSA and XSCORE to compute binding free energy in XIAP-peptide complexes." Journal of Chemical Information and Modeling **47**(1): 134-142.

Ofran, Y. and B. Rost (2007). "Protein-protein interaction hotspots carved into sequences." Plos Computational Biology **3**(7): 1169-1176.

Okada, T., M. Sugihara, et al. (2004). "The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure." J Mol Biol **342**(2): 571-583.

Olson, M. A., M. Feig, et al. (2008). "Prediction of protein loop conformations using multiscale Modeling methods with physical energy scoring functions." Journal of Computational Chemistry **29**(5): 820-831.

Olson, S. T. and I. Bjork (1994). "Regulation of thrombin activity by antithrombin and heparin." Semin Thromb Hemost **20**(4): 373-409.

Olson, S. T., S. Schedin-Weiss, et al. (2010). "Kinetic evidence that allosteric activation of antithrombin by heparin is mediated by two sequential conformational changes." Archives of Biochemistry and Biophysics **504**(2): 169-176.

Oltersdorf, T., S. W. Elmore, et al. (2005). "An inhibitor of Bcl-2 family proteins induces regression of solid tumours." Nature **435**(7042): 677-681.

Ordonez, A., I. Martinez-Martinez, et al. (2009). "Effect of citrullination on the function and conformation of antithrombin." FEBS J **276**(22): 6763-6772.

Orlando, V. and K. A. Jones (2002). "Wild chromatin: regulation of eukaryotic genes in their natural chromatin context." Genes Dev **16**(16): 2039-2044.

Palczewski, K., T. Kumasaka, et al. (2000). "Crystal structure of rhodopsin: A G protein-coupled receptor." Science **289**(5480): 739-745.

Pauling, L. and R. B. Corey (1951). "Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets." Proc Natl Acad Sci U S A **37**(11): 729-740.

Pearlman, D. A. (2005). "Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase." J Med Chem **48**(24): 7796-7807.

Pedersen-White, J. R., L. P. Chorich, et al. (2008). "The prevalence of intragenic deletions in patients with idiopathic hypogonadotropic hypogonadism and Kallmann syndrome." Mol Hum Reprod **14**(6): 367-370.

Peng, H. P. and A. S. Yang (2007). "Modeling protein loops with knowledge-based prediction of sequence-structure alignment." Bioinformatics **23**(21): 2836-2842.

Peng, J. and J. Xu (2010). "Low-homology protein threading." Bioinformatics **26**(12): 294-300.

Perez-Sanchez, H. and W. Wenzel (2011). "Optimization Methods for Virtual Screening on Novel Computational Architectures." Current Computer-Aided Drug Design **7**(1): 44-52.

Petrey, D., Z. X. Xiang, et al. (2003). "Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling." Proteins-Structure Function and Genetics **53**(6): 430-435.

Pettitt, C. S., L. J. McGuffin, et al. (2005). "Improving sequence-based fold recognition by using 3D model quality assessment." Bioinformatics **21**(17): 3509-3515.

Pitteloud, N., J. S. Acierno, Jr., et al. (2006). "Mutations in fibroblast growth factor receptor 1 cause both Kallmann syndrome and normosmic idiopathic hypogonadotropic hypogonadism." Proc Natl Acad Sci U S A **103**(16): 6281-6286.

Pitteloud, N., R. Quinton, et al. (2007). "Digenic mutations account for variable phenotypes in idiopathic hypogonadotropic hypogonadism." J Clin Invest **117**(2): 457-463.

Ponath, P. D., S. Qin, et al. (1996). "Molecular cloning and characterization of a human eotaxin receptor expressed selectively on eosinophils." J Exp Med **183**(6): 2437-2448.

Porollo, A. and J. Meller (2007). "Prediction-based fingerprints of protein-protein interactions." Proteins **66**(3): 630-645.

Qiu, J., W. Sheffler, et al. (2008). "Ranking predicted protein structures with support vector regression." Proteins-Structure Function and Bioinformatics **71**(3): 1175-1182.

Rajamani, D., S. Thiel, et al. (2004). "Anchor residues in protein-protein interactions." Proceedings of the National Academy of Sciences of the United States of America **101**(31): 11287-11292.

Ramachandran, G. N., C. Ramakrishnan, et al. (1963). "Stereochemistry of polypeptide chain configurations." J Mol Biol **7**: 95-99.

Randall, A. and P. Baldi (2008). "SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs." Bmc Structural Biology **8**: -.

Rankin, S. M., D. M. Conroy, et al. (2000). "Eotaxin and eosinophil recruitment: implications for human disease." Molecular Medicine Today **6**(1): 20-27.

Rarey, M., B. Kramer, et al. (1996). "A fast flexible docking method using an incremental construction algorithm." J Mol Biol **261**(3): 470-489.

Rasmussen, S. G., H. J. Choi, et al. (2007). "Crystal structure of the human beta2 adrenergic G-protein-coupled receptor." Nature **450**(7168): 383-387.

Richards, F. M. and C. E. Kundrot (1988). "Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure." Proteins **3**(2): 71-84.

Rinaldi-Carmona, M., F. Barth, et al. (1994). "SR141716A, a potent and selective antagonist of the brain cannabinoid receptor." FEBS Lett **350**(2-3): 240-244.

Roach, J. C., C. Boysen, et al. (1995). "Pairwise end sequencing: a unified approach to genomic mapping and sequencing." Genomics **26**(2): 345-353.

Rogers, J. (2000). "Sequencing the human genome." Journal of Medical Genetics **37**: S30-S30.

Rognan, D. (2007). "Chemogenomic approaches to rational drug design." Br J Pharmacol **152**(1): 38-52.

Rose, P. W., B. Beran, et al. (2011). "The RCSB Protein Data Bank: redesigned web site and web services." Nucleic Acids Research **39**: D392-D401.

Rot, A. (1993). "Neutrophil attractant/activation protein-1 (interleukin-8) induces in vitro neutrophil migration by haptotactic mechanism." Eur J Immunol **23**(1): 303-306.

Roukos, D. H. (2009). "Personal genomics and genome-wide association studies: novel discoveries but limitations for practical personalized medicine." Ann Surg Oncol **16**(3): 772-773.

Rowen, L., G. Mahairas, et al. (1997). "Sequencing the human genome." Science **278**(5338): 605-607.

Sali, A. (1995). "Comparative protein modeling by satisfaction of spatial restraints." Molecular Medicine Today **1**(6): 270-277.

Sali, A. and T. L. Blundell (1993). "Comparative Protein Modeling by Satisfaction of Spatial Restraints." Journal of Molecular Biology **234**(3): 779-815.

Sandor, M., R. Kiss, et al. (2010). "Virtual fragment docking by Glide: a validation study on 190 protein-fragment complexes." Journal of Chemical Information and Modeling **50**(6): 1165-1172.

Sattler, M., H. Liang, et al. (1997). "Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis." Science **275**(5302): 983-986.

Schug, A. and W. Wenzel (2004). "Predictive in silico all-atom folding of a four-helix protein with a free-energy model." J Am Chem Soc **126**(51): 16736-16737.

Schug, A. and W. Wenzel (2006). "An evolutionary strategy for all-atom folding of the 60-amino-acid bacterial ribosomal protein l20." Biophys J **90**(12): 4273-4280.

Schwede, T., J. Kopp, et al. (2003). "SWISS-MODEL: An automated protein homology-modeling server." Nucleic Acids Res **31**(13): 3381-3385.

Schymkowitz, J., J. Borg, et al. (2005). "The FoldX web server: an online force field." Nucleic Acids Research **33**: W382-W388.

Seminara, S. B., S. Messager, et al. (2003). "The GPR54 gene as a regulator of puberty." N Engl J Med **349**(17): 1614-1627.

Sham, Y. Y., Z. T. Chu, et al. (2000). "Examining methods for calculations of binding free energies: LRA, LIE, PDLD-LRA, and PDLD/S-LRA calculations of ligands binding to an HIV protease." Proteins **39**(4): 393-407.

Sharp, K. A., A. Nicholls, et al. (1991). "Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models." Biochemistry **30**(40): 9686-9697.

Shi, J., T. L. Blundell, et al. (2001). "FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties." J Mol Biol **310**(1): 243-257.

Shi, X. F., S. Liu, et al. (2002). "Structural analysis of human CCR2b and primate CCR2b by molecular modeling and molecular dynamics simulation." J Mol Model **8**(7): 217-222.

Shim, J. Y., W. J. Welsh, et al. (2003). "Homology model of the CB1 cannabinoid receptor: sites critical for nonclassical cannabinoid agonist interaction." Biopolymers **71**(2): 169-189.

Simons, K. T., C. Kooperberg, et al. (1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." Journal of Molecular Biology **268**(1): 209-225.

Simonson, T., G. Archontis, et al. (2002). "Free energy simulations come of age: Protein-ligand recognition." Accounts of Chemical Research **35**(6): 430-437.

Singh, V. and P. Somvanshi (2009). "Homology modeling of adenosine A2A receptor and molecular docking for exploration of appropriate potent antagonists for treatment of Parkinson's disease." Curr Aging Sci **2**(2): 127-134.

Sipple, J. H. (1984). "Multiple endocrine neoplasia type 2 syndromes: historical perspectives." Henry Ford Hosp Med J **32**(4): 219-221.

Skelton, N. J., C. Quan, et al. (1999). "Structure of a CXC chemokine-receptor fragment in complex with interleukin-8." Structure **7**(2): 157-168.

Skolnick, J., H. Y. Zhou, et al. (2009). "Performance of the Pro-sp3-TASSER server in CASP8." Proteins-Structure Function and Bioinformatics **77**: 123-127.

Smith, G. R. and M. J. Sternberg (2002). "Prediction of protein-protein interactions by docking methods." Curr Opin Struct Biol **12**(1): 28-35.

Snoep, J. L., F. Bruggeman, et al. (2006). "Towards building the silicon cell: a modular approach." Biosystems **83**(2-3): 207-216.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison (vol 21, pg 951, 2005)." Bioinformatics **21**(9): 2144-2144.

Soto, C. S., M. Fasnacht, et al. (2008). "Loop modeling: Sampling, filtering, and scoring." Proteins **70**(3): 834-843.

Spanagel, R. and F. Weiss (1999). "The dopamine hypothesis of reward: past and current status." Trends Neurosci **22**(11): 521-527.

Spassov, V. Z., P. K. Flook, et al. (2008). "LOOPER: a molecular mechanics-based algorithm for protein loop prediction." Protein Engineering Design & Selection **21**(2): 91-100.

Springer, T. A. (1994). "Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm." Cell **76**(2): 301-314.

Steger, D. J., E. S. Haswell, et al. (2003). "Regulation of chromatin remodeling by inositol polyphosphates." Science **299**(5603): 114-116.

Sternberg, M. J., P. Aloy, et al. (1998). "A computational system for modelling flexible protein-protein and protein-DNA docking." Proc Int Conf Intell Syst Mol Biol **6**: 183-192.

Stoddard, B. L. and D. E. Koshland, Jr. (1992). "Prediction of the structure of a receptor-protein complex using a binary docking method." Nature **358**(6389): 774-776.

Stratikos, E. and P. G. Gettins (1999). "Formation of the covalent serpin-proteinase complex involves translocation of the proteinase by more than 70 A and full insertion of the reactive center loop into beta-sheet A." Proc Natl Acad Sci U S A **96**(9): 4808-4813.

Tanaka, Y., D. H. Adams, et al. (1993). "Proteoglycans on endothelial cells present adhesion-inducing cytokines to leukocytes." Immunol Today **14**(3): 111-115.

Tang, M. L. and C. V. Powell (2001). "Childhood asthma as an allergic disease: rationale for the development of future treatment." Eur J Pediatr **160**(12): 696-704.

Taylor, R. D., P. J. Jewsbury, et al. (2002). "A review of protein-small molecule docking methods." J Comput Aided Mol Des **16**(3): 151-166.

Thoma, N. H., B. K. Czyzewski, et al. (2005). "Structure of the SWI2/SNF2 chromatin-remodeling domain of eukaryotic Rad54." Nat Struct Mol Biol **12**(4): 350-356.

Thompson, J. D., D. G. Higgins, et al. (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): 4673-4680.

Thorn, K. S. and A. A. Bogan (2001). "ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions." Bioinformatics **17**(3): 284-285.

Tucci, S. A., E. K. Rogers, et al. (2004). "The cannabinoid CB1 receptor antagonist SR141716 blocks the orexigenic effects of intrahypothalamic ghrelin." Br J Pharmacol **143**(5): 520-523.

Tuccinardi, T., M. G. Cascio, et al. (2007). "Structure-based virtual screening: Identification of novel CB2 receptor ligands." Letters in Drug Design & Discovery **4**(1): 15-19.

Tuccinardi, T., P. L. Ferrarini, et al. (2006). "Cannabinoid CB2/CB1 selectivity. Receptor modeling and automated docking analysis." J Med Chem **49**(3): 984-994.

Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature **403**(6770): 623-627.

Utgaard, J. O., F. L. Jahnsen, et al. (1998). "Rapid secretion of prestored interleukin 8 from Weibel-Palade bodies of microvascular endothelial cells." J Exp Med **188**(9): 1751-1756.

Vakser, I. A. (1995). "Protein docking for low-resolution structures." Protein Engineering **8**(4): 371-377.

Van Laere, K., C. Casteels, et al. (2010). "Widespread decrease of type 1 cannabinoid receptor availability in Huntington disease in vivo." J Nucl Med **51**(9): 1413-1417.

VanDemark, A. P., M. M. Kasten, et al. (2007). "Autoregulation of the rsc4 tandem bromodomain by gcn5 acetylation." Mol Cell **27**(5): 817-828.

Venter, J. C. (2010). "Multiple personal genomes await." Nature **464**(7289): 676-677.

Venter, J. C. (2011). "Genome-sequencing anniversary. The human genome at 10: successes and challenges." Science **331**(6017): 546-547.

Verma, A., S. M. Gopal, et al. (2007). "All-atom de novo protein folding with a scalable evolutionary algorithm." J Comput Chem **28**(16): 2552-2558.

Verma, A., S. M. Gopal, et al. (2008). "All-atom protein folding with free-energy forcefields." Prog Mol Biol Transl Sci **83**: 181-253.

Verma, A. and W. Wenzel (2009). "A free-energy approach for all-atom protein simulation." Biophys J **96**(9): 3483-3494.

Verma, H., P. R. Patil, et al. (2008). "Antiviral activity of the Indian medicinal plant extract Swertia chirata against herpes simplex viruses: a study by in-vitro and molecular approach." Indian J Med Microbiol **26**(4): 322-326.

Verty, A. N., J. R. McFarlane, et al. (2004). "Evidence for an interaction between CB1 cannabinoid and melanocortin MCR-4 receptors in regulating food intake." Endocrinology **145**(7): 3224-3231.

Vilar, S., G. Cozza, et al. (2008). "Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery." Current Topics in Medicinal Chemistry **8**(18): 1555-1572.

Vissers, L. E., C. M. van Ravenswaaij, et al. (2004). "Mutations in a new member of the chromodomain gene family cause CHARGE syndrome." Nat Genet **36**(9): 955-957.

Voegtli, W. C., A. Y. Madrona, et al. (2003). "The structure of Aip1p, a WD repeat protein that regulates Cofilin-mediated actin depolymerization." J Biol Chem **278**(36): 34373-34379.

Wallner, B. and A. Elofsson (2006). "Identification of correct regions in protein models using structural, alignment, and consensus information." Protein Science **15**(4): 900-913.

Wallner, B., H. S. Fang, et al. (2003). "Automatic consensus-based fold recognition using Pcons, ProQ, and pmodeller." Proteins-Structure Function and Genetics **53**(6): 534-541.

Wallner, B., P. Larsson, et al. (2007). "Pcons.net: protein structure prediction meta server." Nucleic Acids Res **35**(Web Server issue): W369-374.

Wang, R., X. Fang, et al. (2004). "The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures." J Med Chem **47**(12): 2977-2980.

Wang, R., Y. Lu, et al. (2003). "Comparative evaluation of 11 scoring functions for molecular docking." J Med Chem **46**(12): 2287-2303.

Wang, Z., A. N. Tegge, et al. (2009). "Evaluating the absolute quality of a single protein model using structural features and support vector machines." Proteins-Structure Function and Bioinformatics **75**(3): 638-647.

Warkentin, T. E., B. H. Chong, et al. (1998). "Heparin-induced thrombocytopenia: towards consensus." Thromb Haemost **79**(1): 1-7.

Warren, G. L., C. W. Andrews, et al. (2006). "A critical assessment of docking programs and scoring functions." J Med Chem **49**(20): 5912-5931.

Warren, G. L., W. Andrews, et al. (2004). "Critical assessment of docking programs and scoring functions." Abstracts of Papers of the American Chemical Society **228**: U513-U514.

Wei, B. Q., L. H. Weaver, et al. (2004). "Testing a flexible-receptor docking algorithm in a model binding site." J Mol Biol **337**(5): 1161-1182.

Wells, J. A. and C. L. McClendon (2007). "Reaching for high-hanging fruit in drug discovery at protein-protein interfaces." Nature **450**(7172): 1001-1009.

Wenzel, W., B. Fischer, et al. (2007). "Accuracy of binding mode prediction with a cascadic stochastic tunneling method." Proteins-Structure Function and Bioinformatics **68**(1): 195-204.

Wenzel, W., B. Fischer, et al. (2008). "Receptor-specific scoring functions derived from quantum chemical models improve affinity estimates for in-silico drug discovery." Proteins-Structure Function and Bioinformatics **70**(4): 1264-1273.

Wenzel, W. and K. Hamacher (1999). "Stochastic tunneling approach for global minimization of complex potential energy landscapes." Physical Review Letters **82**(15): 3003-3007.

Wenzel, W. G. and D. B. Kokh (2008). "Flexible side chain models improve enrichment rates in in silico screening." Journal of Medicinal Chemistry **51**(19): 5919-5931.

Wiederstein, M. and M. J. Sippl (2007). "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins." Nucleic Acids Research **35**: W407-W410.

Wiehe, K., M. W. Peterson, et al. (2008). "Protein-protein docking: overview and performance analysis." Methods Mol Biol **413**: 283-314.

Wu, C. H., R. Apweiler, et al. (2006). "The Universal Protein Resource (UniProt): an expanding universe of protein information." Nucleic Acids Res **34**(Database issue): D187-191.

Xiang, Z. (2006). "Advances in homology protein structure modeling." Curr Protein Pept Sci **7**(3): 217-227.

Xiang, Z., C. S. Soto, et al. (2002). "Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction." Proc Natl Acad Sci U S A **99**(11): 7432-7437.

Xu, N., H. G. Kim, et al. (2011). "Nasal embryonic LHRH factor (NELF) mutations in patients with normosmic hypogonadotropic hypogonadism and Kallmann syndrome." Fertil Steril **95**(5): 1613-1620 e1611-1617.

Yan, A., A. Kloczkowski, et al. (2007). "Prediction of side chain orientations in proteins by statistical machine learning methods." J Biomol Struct Dyn **25**(3): 275-288.

Yanover, C., O. Schueler-Furman, et al. (2008). "Minimizing and learning energy functions for side-chain prediction." Journal of Computational Biology **15**(7): 899-911.

Yatsenko, S. A., S. W. Cheung, et al. (2005). "Deletion 9q34.3 syndrome: genotype-phenotype correlations and an extended deletion in a patient with features of Opitz C trigonocephaly." J Med Genet **42**(4): 328-335.

Yogurtcu, O. N., S. B. Erdemli, et al. (2008). "Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations." Biophysical Journal **94**(9): 3475-3485.

Zaki, M. J. and C. Bystroff (2008). Protein structure prediction Texte imprimÂe. Totowa (N.J.), Humana press.

Zhang, W. and Y. Duan (2006). "Grow to Fit Molecular Dynamics (G2FMD): an ab initio method for protein side-chain assignment and refinement." Protein Engineering Design & Selection **19**(2): 55-65.

Zhang, Y. (2007). "Template-based modeling and free modeling by I-TASSER in CASP7." Proteins **69 Suppl 8**: 108-117.

Zhang, Y. (2008). "I-TASSER server for protein 3D structure prediction." BMC Bioinformatics **9**: 40.

Zhang, Y. (2008). "Progress and challenges in protein structure prediction." Current Opinion in Structural Biology **18**(3): 342-348.

Zheng, X. F. and T. F. Chan (2002). "Chemical genomics: a systematic approach in biological research and drug discovery." Curr Issues Mol Biol **4**(2): 33-43.

Zhu, K., D. L. Pincus, et al. (2006). "Long loop prediction using the protein local optimization program." Proteins-Structure Function and Bioinformatics **65**(2): 438-452.

Zwanzig, R. W. (1954). "High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases." J. Chem. Phys. **22**: 1420.

Acknowledgements

First of all I especially want to thank my advisor Apl. Prof. Wolfgang Wenzel: thanks for your energy and enthusiasm that motivated me in my PhD and for inspiring me in continuing working in the research field.

Special thank to my mum and dad for their support, without their help and encouragement, this thesis would have not been possible; and to Fabio, for motivating me, for his good advice and for beeing there always when I need it.

I would like to thank Horacio for his friendship and help in the past three years, from you I have just not received help, but I have learned a lot, thanks; and to Timo who was always there to help me when I needed (I would have never started without your help).

Thank also to my colleague Konstantin, it was a pleasure to be in the office with him, to Simon for his kind help in the Zuzammenfassung, and to Priya for the good advice with the references and of course to all the other people of the group and colleagues with whom I have cooperated.

In addition I would like to thank Prof. Dr. Stefan Bräse, Dr.Katja Schmitz, Hyung Goo Kim and Prof. Layman for their interesting and successful collaborations.