AUTOMATED IN SILICO PROTEIN MODELING STRATEGIES - APPLICATIONS AND LIMITS IN G-PROTEIN COUPLED RECEPTOR MODELING

Dissertation

zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
der Naturwissenschaftlich-Technischen Fakultät I
- Mathematik und Informatik -
der Universität des Saarlandes

vorgelegt von

Benny Kneissl

Universität des Saarlandes
Saarbrücken, 2012

Tag des Kolloquiums	30.11.2012
Dekan	Univ.-Prof. Mark Groves
Berichterstatter	Prof. Dr. Andreas Hildebrandt
	Prof. Dr. Hans-Peter Lenhof
Vorsitz	Prof. Dr. Bläser
Akad. Mitarbeiter	Dr. Marc Hellmuth

Eidesstattliche Versicherung

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

Saarbrücken, 2012

Benny Kneissl

ABSTRACT

In the 21st century, computer-aided methods became a well-established tool in the pharmaceutical industry. Because many protein structures are still very difficult to determine experimentally, various in silico methods have been developed to predict their three-dimensional structure from the corresponding sequence. Unfortunately, the modeling process has often to be adapted to the target protein or the automatically generated models have to be manually refined afterwards by the users based on their expert knowledge to finally achieve reasonable results.
The aim of this thesis is to explore the applicability of automated in silico modeling strategies by means of G-protein coupled receptors (GPCRs). First, we analyze to which extent available protein structure prediction methods can be particularly tailored to the automated GPCR modeling case. Second, we develop our own approach and demonstrate that we obtain improved models compared to other state-of-the-art modeling tools. More important, our method does not rely on manual interactions by the user during the modeling process and is thus generally applicable for all GPCRs. Furthermore, we present a new sequence based method to predict structural distortions from ideal α-helical geometry. This method also exceeds the prediction accuracy of comparable approaches.

GERMAN ABSTRACT

Im 21. Jahrhundert sind computergestützte Methoden zu einem etablierten Werkzeug in der Pharmaindustrie geworden. Da viele Proteinstrukturen nur mit großem Aufwand experimentell aufgeklärt werden können, wurden diverse in silico Methoden entwickelt, um ihre dreidimensionale Struktur aus der dazugehörigen Sequenz zu modellieren. Leider muss der Modellierungsprozess häufig an das Zielprotein angepasst oder die automatisch generierten Modelle von den Anwendern auf Basis ihres Fachwissens verfeinert werden, um letztendlich vernünftige Resultate zu erhalten.
Das Ziel dieser Arbeit ist die Untersuchung der Anwendbarkeit automatischer in silico Modellierungsstrategien am Beispiel der G-Protein gekoppelten Rezeptoren (GPCRs). Zunächst analysieren wir bis zu welchem Grad verfügbare Proteinstruktur-Vorhersagemethoden für die automatische Modellierung von GPCRs angepasst werden können. Anschließend entwickeln wir unser eigenes Verfahren und zeigen, dass dieses im Vergleich zu anderen modernen Methoden verbesserte Ergebnisse erzielt. Hervorzuheben ist dabei, dass unser Ansatz keinerlei Interaktion vom Anwender benötigt und somit auf alle GPCRs angewendet werden kann. Des Weiteren stellen wir eine neue sequenzbasierte Methode zur Vorhersage von so genannten Kinks in α-Helices vor. Auch diese Methode übertrifft die Vorhersagegenauigkeit vergleichbarer Ansätze.

GERMAN SUMMARY

Im 21. Jahrhundert sind computergestützte Methoden zu einem etablierten Werkzeug in der Pharmaindustrie geworden. Zum einen dienen zum Beispiel Methoden aus dem maschinellem Lernen der automatischen Analyse riesiger Datenmengen, zum anderen können Modellierungsmethoden bei der Vorhersage von Proteinstrukturen helfen und schließlich zur Entdeckung neuer Wirkstoffe führen. Mit Hilfe dieser Verfahren sollen in Zukunft Teile der kosten- und zeitintensiven Laborexperimente sogar komplett ersetzt werden können.

Alle in silico entwickelten Strukturvorhersagemethoden werden grob in zwei Klassen eingeteilt. Die erste ist die sogenannte *Homologiemodellierung*, bei dem die noch unbekannte Struktur eines Proteins aus bekannten Strukturen abgeleitet wird. Die Basis dieses Verfahrens ist die Erkenntnis, dass im Gegensatz zur Primärstruktur, das heißt der Aminosäuresequenz des Proteins, die Tertiärstruktur deutlich konservierter ist. Aus diesem Grund findet dieser Ansatz seine häufige Anwendung, wenn homologe Proteine, z.B. aus der selben Proteinfamilie, als Templat zur Verfügung stehen.
Dieser Methode steht die *ab initio Modellierung* gegenüber, die meist angewendet wird, wenn noch kein passendes Templat experimentell bestimmt worden ist. Somit muss die komplette strukturelle Information allein aus der zu Grunde liegenden Sequenz extrahiert bzw. vorhergesagt werden. Der erste Schritt ist zunächst die Bestimmung aller Sekundärstrukturelemente und anschließend die Anordnung dieser im dreidimensionalen Raum. Die so generierten Decoys werden daraufhin bezüglich einer Energiefunktion optimiert.
Wie häufig bei der Anwendung automatischer Methoden auf biologische Daten muss auch hier fast jeder Schritt auf Basis von Expertenwissen manuell kontrolliert und korrigiert werden.

In dieser Arbeit untersuchen wir deshalb, in wieweit die oben genannten Methoden zur Strukturvorhersage von Proteinen aus der Familie der G-Protein gekoppelten Rezeptoren (GPCRs) verwendet werden können ohne in die Modellierung eingreifen oder nachträglich die generierten Modelle manuell modifizieren zu müssen. Wegen ihrer biologischen Funktionsvielfalt, unter anderem der Regulierung des Blutdrucks oder der Immunsystemaktivität, sowie als Ursache von vielen gängigen und ernsthaften Krankheiten (Diabetis, Alzheimer, Parkinson) aufgrund einer Fehlfunktion, sind GPCRs seit Jahren im Fokus der pharmazeutischen Industrie. Da diese Proteine in der Zellmembran sitzen, sind experimentelle Methoden zur Aufklärung der Struktur, wie z.B. die Röntgenkristallographie, kaum anwendbar, so dass lediglich die Struktur eines Vertreters dieser Familie, bovine rhodopsin (PDB ID: 1F88), zu Beginn dieser Arbeit bekannt war. Deshalb ist es besonders wichtig, die generelle Anwendbarkeit von in silico Methoden für diese Proteinfamilie zu analysieren.

In unserer ersten Studie untersuchen wir die Modellierung eines noch unbekannten GPCRs, dem Neurokinin-1 Rezeptor, mittels Homologiemodellierung. Im Gegensatz zu anderen Studien verwenden wir jedoch zwei Template, bovine rhodopsin und den zwischenzeitlich experimentell aufgelösten β_1-adrenergic Rezeptor (PDB ID: 2RH1), gleichzeitig. Wir zeigen, dass der dadurch erhöhte Konformationsraum des Proteinrückgrats zu einer strukturellen Variation der Modelle führt, so dass automatisch Modelle erzeugt werden, die zu experimentellen Studien passen. Die anschließende Auswahl eines geeigneten Modells kann allerdings bisher nur auf Basis dieser Studien geschehen und somit nicht automatisiert werden.

Das zweite Projekt der zu Grunde liegenden Arbeit ist die Analyse eines ab initio Verfahrens, welches speziell für die Modellierung von GPCRs entwickelt wurde. Da das Programm selbst nicht für die Öffentlichkeit zugänglich ist, implementieren wir das Verfahren zunächst selbst, um die in der Publikation beschriebenen vielversprechenden Ergebnisse nachvollziehen zu können. Neben der reinen Nachimplementierung steht vor allem auch die Verbesserung auf algorithmischer Ebene im Vordergrund. Bei vielen Schritten, insbesondere bei der Generierung tauglicher Startstrukturen, sind die Ergebnisse jedoch unzureichend, so dass die anschließende Optimierung der Modelle fehlschlägt. Auch diverse Anpassungen unsererseits – ohne den Kern des Algorithmus zu verändern – reichen nicht aus, um mit dieser Methode erfolgreich GPCRs modellieren zu können.

Im Hinblick auf unser Ziel der automatisierten GPCR Modellierung analysieren wir die zwischenzeitlich experimentell neu aufgelösten Strukturen von fünf weiteren GPCRs. Bei dem Vergleich wird deutlich, dass die strukturellen Hauptunterschiede der GPCRs in der Lage und Orientierung der Kinks bestehen. Kinks sind Abweichungen von der idealen Geometrie einer Helix, die nicht nur aufgrund der (lokalen) Aminosäuresequenz, sondern auch aufgrund inter-helikaler Wechselwirkung ausgebildet werden. Gerade bei aus einfachen Sekundärstrukturelementen bestehenden Proteinen dürfen Kinks bei der Modellierung also nicht vernachlässigt werden.

Da bisher die Ursache für solche Kinks noch nicht bis ins Detail geklärt ist und auch nur wenige Vorhersagemethoden existieren, haben wir zunächst eine eigene Vorhersagemethode entwickelt. Mittels Support-Vektor-Maschinen können wir – besser als vergleichbare Ansätze – gekinkte Helizes mit über 80%iger Genauigkeit vorhersagen. Dennoch bleibt die Vorhersage der exakten Lage des Kinks noch ungelöst, was darauf zurückzuführen ist, dass, wie bereits erwähnt, inter-helikale Wechselwirkungen eine wichtige Rolle spielen.

Basierend auf der Ähnlichkeit der bekannten GPCRs haben wir unser eigenes Verfahren zur automatischen Modellierung dieser entwickelt. Unser Ansatz verknüpft dabei die Vorteile der Homologiemodellierung für die Generierung geeigneter Startstrukturen, kommt jedoch ohne Sequenzalignment aus, welches einer der fehleranfälligsten Schritte bei der Homologiemodellierung ist. Lediglich die am stärksten konservierte Aminosäure in jeder Helix muss identifiziert werden. Des Weiteren erlaubt unser Verfahren den Helices die Ausbildung der Kinks während der Optimierungsprozedur. Somit werden auch inter-helikale Wechselwirkungen

bei der Optimierung der Kinks in Betracht gezogen. Durch dieses in der Form bisher nie dagewesene Verfahren können alle bekannten GPCRs in der Transmembranregion zumeist mit einem C_α-RMSD Wert unter 2.0Å und mit einem maximalen Wert von 2.65Å modelliert werden. Damit erhalten wir verbesserte Modelle im Vergleich zu anderen gängigen Methoden aus der Literatur.

ACKNOWLEDGEMENTS

Although only my name appears on the cover of this thesis, many people have contributed to its production. I would like to thank everybody but I have to express my apology that I cannot mention personally one by one.

First and foremost, I want to thank my supervisor Prof. Dr. Andreas Hildebrandt. He asked excellent inspiring questions and it often took him less than 5 minutes to re-boost my research. When commenting on my views he helped me to understand and to enrich my own ideas. I am grateful to him for holding me to a high research standard, and thus teaching me how to do research. Moreover, his group has been a source of friendships as well as good advice and collaboration.

I was extraordinarily fortunate to work for the company Boehringer Ingelheim. Therefore, thanks to Dr. Herbert Köppen for giving me the opportunity of this cooperation. Boehringer Ingelheim funded to a large extent my PhD thesis but it was even more important to have the chance to closely work with computational biologists from pharmaceutical industry. Here, I would particularly like to mention Dr. Christofer Tautermann, whose extensive biological background and experiences in protein modeling have always been a key element in solving my tasks. I am also thankful to him for carefully reading and commenting on countless revisions of my manuscripts.

I also want to acknowledge Prof. Dr. Hans-Peter Lenhof, who supported me from the beginning of my studies. At his chair, I did my first research projects as a student researcher and in terms of my bachelor's and master's thesis. Furthermore, he gave me his confidence to be a tutor of his lectures, which have been my first experiences in teaching. I am indebted to the members of his chair for shared projects and private activities, too.

It is a pleasure to pay tribute to the collaborators in my GPCR modeling project, Sophie Weggler, Kristyna Pluhackova, Alexander Rurainski and, in particular, Sabine Müller, who has done a great deal of groundwork in the kink prediction project. I think they know from their own experiences how important supporting colleagues are, and hence how thankful I am. I would also like to express my thanks to René Hussong, Stefan Nickels, Anne Dehof, Daniel Stöckel, and Nina Fischer who have been collaborators in the BALL project as well as my new colleagues, Marco Carnini and Markus Krupp, who ensured a pleasant working environment.

Many side-projects have been done in cooperation with students in terms of diploma, bachelor's or master's thesis. Therefore, thanks to Thomas Thies, Jan Riehm, Alexander Baldauf, Debora Ernst, Andreas Lund and Tim Seifert.

I also want to thank my two fellow students Matthias Dietzen and Andreas Keller, who had a significant influence on the successful completion of my studies. In many night sessions we discussed the current lectures, solved the exercises or learned for the exams such that we all finished our studies faster and better than expected. Over about 10 years both are rather close friends than just colleagues.

Many friends have helped me stay sane through the last years. Due to their support, I was able to overcome setbacks and stay focussed on my graduate studies and PhD. It is impossible to name all friends but I greatly value the friendship with Lars Steinbrück, Christina Backes, Sebastian Kirschbaum and Daniel Dumitriu.

Most importantly, none of this would have been possible without the love and patience of Bettina Leonhardt. The large distance in our relationship made things sometimes difficult but her mental support was always the foundation of my work. I deeply appreciate her belief in me through the last years and also her courage to go on all adventures I have in mind.

I also would like to express my heart-felt gratitude to my family, in particular my parents. They have been a constant source of love, concern, support and strength all these years.

I want to end my thesis acknowledgement by thanking the person who 'discovered' coffee. Everyone who has a chance to look at one of our institute's coffee lists knows why. ☺

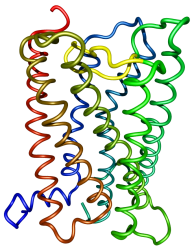
CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADMET	Pharmacology for Absorption, Distribution, Metabolism, Excretion, and Toxicity
AE	Artificial Evolution
DoF(s)	Degree(s) of Freedom
EC	Extracellular
ECL	Extracellular Loop
GPCRs	G-Protein Coupled Receptors
H1 ... H7	Helix 1 ... Helix 7
HTS	High-Throughput Screening
IC	Intracellular
MD	Molecular dynamics
NMR	Nucleic Magnetic Resonance
PCA	Principal Component Analysis
PDB	Protein Data Base
SA	Simulated Annealing
SASA	Solvent Accessible Surface Area
SCO	Side Chain Optimization
SCWRL	Side Chain With Rotamer Library
SSP	Secondary Structure Prediction
TFE	Trifluoroethanol



1 *Introduction*

Computer-aided methods are increasingly applied in many areas of pharmaceutical industry. Today, scientists can choose from a large number of databases to access and to handle the huge amount of available experimental data, they can employ various machine learning methods for data analysis, and can use many sophisticated tools that assist in structure prediction of unknown proteins.

The latter methods are especially important due to the limits of experimental techniques, of which *X-ray crystallography* is certainly the most common one. In 1958, Max Perutz and John Kendrew, who have been honored with the Nobel Prize in Chemistry, used this method to resolve the first protein structure, myoglobin.[1] In June 2012, the time of writing this thesis, 67618 structures were determined through X-ray crystallography and are deposited in the Protein Data Base (PDB),[2] accounting for nearly 90% (76383) of all experimentally solved protein structures. However, this method is hardly applicable for membrane embedded proteins. In the case of membrane protein crystallization, the protein has to be removed from its natural environment, a phospholipid bilayer, and consequently the protein often changes its conformation. None of the stabilizing procedures guarantees that the structures remain unchanged, and thus crystal packing artefacts cannot be excluded. The crystallization step is further complicated by the hydrophobic surface of these proteins. Today, only 342 structures of membrane proteins are known.[3]
Besides X-ray crystallography, *Nucleic Magnetic Resonance* (NMR) is the second most common technique, which provides atomic resolution of protein structures. At the time of writing, 8286 protein structures were resolved by NMR, where 88 unique proteins belong to the family of integral membrane proteins. The first structure solved by NMR was the bull seminal proteinase inhibitor (BUSI) of a globular protein in 1985.[4] The primary disadvantage of NMR are the difficulties in the 3D structure determination of larger proteins (>250 residues) due to huge overlaps in its corresponding spectrum.[5]

In this study, we focus on G-protein coupled receptors (GPCRs), the largest family of transmembrane proteins. Because each GPCR consists of more than 300 residues, all structures have to be determined by X-ray crystallography. One of the major challenges in GPCR crystallography is the low expression level in native tissues.[6] Moreover, due to the poor thermal stability[7] many different (structural) modifications are required to obtain crystal structures of GPCRs: the addition of lipids during purification and crystallization, the usage of stabilizing ligands and mutations, and the insertion of T4 lysozyme in the disordered region between helix 5 and helix 6 are the most common ones.[8–11] The latter brought the biggest engineering progress in stabilizing GPCRs. Due to this modification, the Stevens &

Kobilka group were able to successfully determine and publish five new GPCRs in 2012.[12–16] The basis of our work and all conclusions, however, are drawn from the native structures available at the end of 2011.

In December 2011, only 7 crystal structures of 799 sequences were experimentally solved. In the GRAFS (**G**lutamate - **R**hodopsin - **A**dhesion - **F**rizzled/Taste2 - **S**ecretin) system[17] - a system to classify the sequences of GPCRs in different subfamilies -, all these structures belong to the rhodopsin-like subfamily (see Figure 1.1). Rhodopsin itself was the first GPCR that was experimentally resolved in 2000 (PDB ID: 1F88), while the others have been determined between 2007 and 2011.

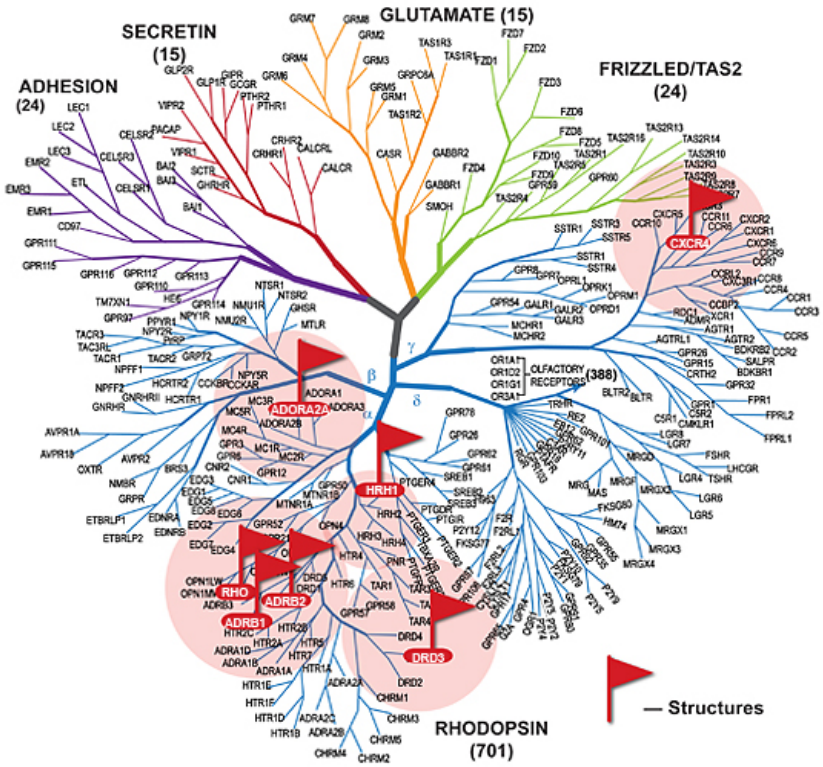


Figure 1.1: The GPCR tree of the five subclasses. All already resolved crystal structures (red flags) belong to the rhodopsin-like subclass and are evolutionary closely related except for the CXCR4 chemokine receptor. Image by Yekaterina Kadyshevskaya, courtesy of the GPCR Network, The Scripps Research Institute.[18]

In general, the function of this protein family is the signal transduction from the outside to the interior of cells. They are thereby involved in various biological processes, e. g., regulation of blood pressure, heart rate, and immune system activity, and hence they are of great interest for the pharmaceutical industry.[19,20] This interest is further increased since dysfunctions of GPCRs can lead to serious diseases such as asthma[21] or schizophrenia.[22] Drugs developed to target GPCRs do not need to have special properties regarding the ability to pass cell membranes. Instead of passing small molecules (drugs) through the cell membranes, GPCRs provide an

outer membrane binding site. The binding of a molecule induces then a conformational change in the intracellular (IC) side. Upon exchange of GDP to GTP the bound G-protein is then dissociated and transfers the signal to various effectors. More information about GPCRs' predominance as drug targets and some details of each solved structure are briefly described in Chapter 2.

A common technique to find a molecule that binds at a specific target, is *High-Throughput Screening* (HTS). Through this cost and time consuming process, millions of compounds can automatically be biologically screened to identify so-called *hits*. However, hits are still far away to be a potential drug and can only be used as a starting point for a mostly very long process of further improvements. Here, computer-aided methods, in particular, *ligand-based* and *structure-based drug design*, are normally used to find new compounds and to improve their profile, i. e., activity, selectivity, and pharmacological properties (ADMET). The idea behind the ligand-based strategy is that a protein can only bind to very similar ligands and hence new ligands can be derived from a known one. In contrast, in structure-based drug design, the ligands are developed based on the binding pocket of a given target protein.

As on the one hand, structural information for GPCRs is still limited, but on the other hand, the demand for additional structures is exceedingly high, we explore the applicability of in silico modeling strategies on this protein family. Two main families of methods are known for this task, *homology modeling* and *ab initio modeling*. Homology modeling, also known as *comparative modeling*, is based on the idea that structures of proteins are more conserved than their sequence, and hence, in case of high sequence identity (commonly $\geqslant 40\%$), a protein with a known three-dimensional structure can serve as a template for a target protein, where only the sequence is known. The general steps in homology modeling are template selection, sequence alignment, structure modeling and model validation. Many algorithms have been implemented in different modeling tools, e.g. fragment assembly (3D-JIGSAW,[23] CPHModel[24]) or satisfaction of spatial restraints (MODELLER,[25] Geno3D[26]).
Many unknown GPCRs have been modeled in silico using homology modeling using rhodopsin as a template since rhodopsin has been the only available structure for several years. However, due to the low identity between GPCR sequences, most of these models had to be manually restrained in the modeling process or refined afterwards based on mutagenesis data and expert knowledge. These restraints had to be adapted for every new target structure researchers focused on and therefore cannot be applied in general. In a recent study, Zhu and Li demonstrated this by modeling the β1-adrenoceptor receptor.[27] Only if they choose the correct alignment and template structure, which they did based on the known target structure, it was possible to model this GPCR with a quite low C_α-RMSD value.

Whereas in comparative modeling an adequate template is needed, for example, a structure of a protein from the same family, in ab initio modeling the structure is predicted solely based on its sequence, e. g., by *protein threading*.

A software tool where protein threading is implemented, called TASSER,[28] has been applied to all identified G-protein coupled receptors.[29] The C_α-RMSD in the core region between their predicted model and the native bovine rhodopsin structure was 3.3 Å. Focusing only on the TM region, the C_α-RMSD was decreased to 2.1 Å.
In a general protein modeling approach, Sander et al. computed the three-dimensional structure of various proteins from evolutionary sequence variation.[30] Whereas some reasonable structures were obtained for other proteins, the representative of the GPCR family, bovine rhodopsin, was only modeled with a C_α-RMSD value of 4.84Å. The authors claimed that the largest differences occurred in helices 1 and 7, which were misaligned relative to the direction perpendicular to the membrane surface.
Algorithms particularly tailored to the GPCR case obtain much better results. One of the first ones is MembStruk. Several version have been announced between 2001 and 2004, until the authors were able to reproduce the transmembrane region of bovine rhodopsin within a C_α-RMSD of 2.8Å.[31]
Already in 2001, Shacham et al. developed PREDICT and modeled bovine rhodopsin with an RMSD of 3.87Å.[32] Excluding helix 4 yielded an even lower RMSD of 3.2Å. But more important, the authors claimed to have reproduced the binding site for retinal (where helix 4 is not involved) very accurately (without giving an RMSD value). It was not until 2004 that the authors published their approach with more details.[33]
The basic idea behind PREDICT is to optimize the arrangement of seven canonical helices and to insert afterwards kinks in each helix using molecular dynamics (MD) simulations without changing the final structure significantly. Because the three-dimensional optimization procedure is a brute force method in a very small range, the whole modeling approach is strongly dependent on proper start conformations and hence on the quality of the prediction methods used for this task. Since the whole procedure of PREDICT - as far as stated in the corresponding publication - did not rely on experimental data and since we were interested in fully automated GPCR modeling, we examined both the reproducibility of their results regarding bovine rhodopsin and the applicability of PREDICT on other GPCRs.

1.1 THE AIM OF THIS THESIS

When this thesis was started in 2006, only one GPCR structure was determined (bovine rhodopsin) and hence GPCR homology modeling was done using this as template. In 2007, a second native GPCR structure, the human β_2-adrenergic receptor, was published and researcher began to use the new available template (PDB ID: 2RH1) for their studies.[34,35] All these studies used restraints to place important side chains in the right position.
Our first goal towards automated GPCR modeling was to examine to which extend manually defined restraints in homology modeling can be replaced by including more flexibility in the backbone. Therefore, we used a multiple template approach to explore whether suitable models can be created without manual influences in the modeling procedure. We only fall back to mutagenesis data to choose a final model,

which is then validated by a virtual screening technique. The details of this study with all results are illustrated and discussed in Chapter 3.

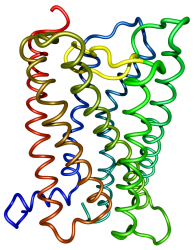
Our second task was to reimplement the PREDICT approach based on the information given in the corresponding publication since the code was closed source. Although PREDICT was successfully applied to rhodopsin, the simple methods used, in particular, the severe restraint of rigid canonical helices, let us infer that it can hardly be applied for automated GPCR modeling. However, we aimed to replace these methods by more sophisticated algorithms to possibly fill this gap.
To this end, we first checked the prediction methods used by PREDICT to find good start conformations, so-called decoys. This analysis includes mainly the orientation of a single helix and the scoring function in 2D. In a second step, we investigated the energy function of PREDICT used in the optimization procedure, which was adapted for the reduced protein representation developed by Herzyk and Hubbard.[36] Here we focused, among others, on the different parameters for cation-pi, polar and aromatic interactions, as well as the given membrane term. Another step of the PREDICT approach was the MD simulation to insert kinks in helices. We analysed, similar to Shacham and coworkers, the introduction of kinks in a helix using an MD simulation. In contrast to their proceeding where they solvated the whole structure in water and ran a very short simulation of 280ps to change the model only locally, we focused on the stability of a single helix and hence we solvated each helix in a mixture of trifluoroethanol (TFE) and water and have run the simulation for two nanoseconds. All investigations concerning the PREDICT algorithm are discussed in Chapter 4.

Besides our primary project – fully automated ab initio GPCR modeling – , we focused on distortions in α-helices (see Chapter 5). As illustrated in Chapter 2, GPCRs consist mainly of seven helices connected by loops of different length. In case of proteins consisting only of such simple building blocks, the largest difference results from distortions from optimal α-helical geometry. As the reasons for those are widely unknown, we extended our analysis in this direction and applied string kernels for support vector machines to predict kinks based on a peptide sequence.

The results obtained in the studies above encouraged us to develop a new optimization method, which allows helices to be kinked during the arrangement of the seven helices. In Chapter 6, the mathematical background and all results concerning this study are presented and discussed.

Chapter 7 concludes our previously presented results. Some final remarks are given and some ideas for future work are suggested.

2 G-Protein Coupled Receptors



G-protein coupled receptors (GPCRs), also known as seven transmembrane receptors (7TM receptors), are the largest family of α-helical transmembrane proteins. Today, more than 700 GPCR-encoding genes are known, starting with the first cloned member, the hamster β-adrenergic receptor, in 1986.[37,38]

Located in the cell membrane, GPCRs have an extracellular and a cytoplasmic binding domain, and are therefore predestined for signal transduction. Since they bind to a wide variety of ligands - from small molecules to proteins -, they clearly belong to the most important pharmaceutical targets. About 50-60% of approved drugs and about 40% of the top selling drugs target a receptor of this family.[39]

Drug	Market value	Target
Statin LIPITOR	$ 13,288	HMG-CoA Reductase
Anti-coagulant PLAVIX	$ 9,100	ADP Receptor - GPCR
Antacid NEXIUM	$ 8,236	H+/K+-ATPase
Asthma treatment SERETIDE	$ 8,099	Adrenoceptor - GPCR
Anti-psychotic SEROQUEL	$ 6,012	Several GPCRs
Anti-inflammatory ENBREL	$ 5,863	TNF receptor
Anti-inflammatory REMICADE	$ 5,453	TNF receptor
Statin CRESTOR	$ 5,383	HMG-CoA Reductase
Anti-psychotic ZYPREXA	$ 5,357	Serotonin Receptor - GPCR
Anti-inflammatory HUMIRA	$ 5,032	TNF receptor

Table 2.1: Top 10 selling products in 2009 (in billion).[39]

The functions of GPCRs are as diverse as their ligands are. They are, among other things, responsible for several automatic body functions such as blood pressure and heart rate,[19] regulation of the immune system activity[20] and digestive processes.[40] Many wide-spread and some serious diseases are related to dysfunctions of these receptors, e.g. hypertension,[19] asthma,[21] schizophrenia,[22] allergic reactions,[41] and Parkinson's disease,[42] to mention only a few.

2.1 SIGNAL TRANSMISSION: G-PROTEIN ACTIVATION

Proteins of the 7TM receptor class all share a common feature – a so-called G-protein bound to their intracellular side. These G-proteins (guanine nucleotide-binding proteins) consist of three subunits ($\alpha\beta\gamma$) and are bound via their α-subunit to an inactive GPCR as shown in Figure 2.1. The activation process, which is not fully understood today, is initiated by the binding of a ligand in the receptor's binding pocket. These ligands are very diverse and can range from ions to whole proteins. The ligand binding causes a conformational change in the intracellular GPCR domain, which in turn results in a detachment of the G-protein from the receptor. In the next step, the GDP in the α-subunit is exchanged with a GTP, leading to a dissociation from the $\beta\gamma$-subunit. Depending on the type of the subunit, various proteins in the cytoplasmic area are then affected. The whole activation process that has the big advantage that molecules do not have to pass the cell membrane is schematically illustrated in Figure 2.2.

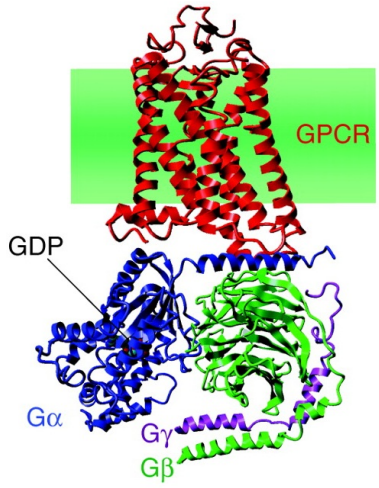


Figure 2.1: GPCR with bound G-protein.[43]

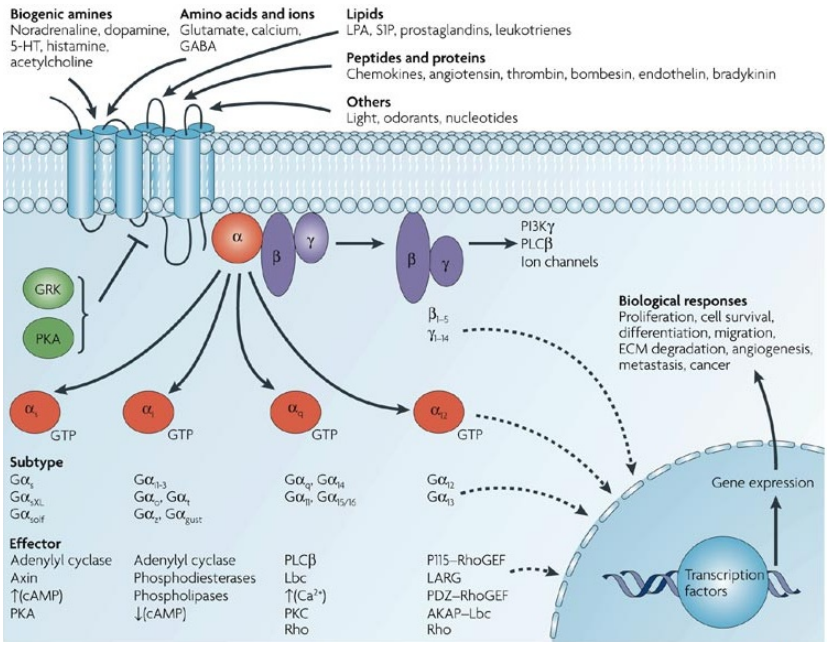


Figure 2.2: Activation process of GPCRs.[44]

2.2 SEQUENCE MOTIFS

According to a study of Illergård et al., protein structures are up to ten times more conserved than their sequences.[45] G-protein coupled receptors are an excellent example for these observations. The sequence similarity within this family is very low (usually $\leqslant 40\%$), although their structures have many properties in common (see Section 2.3). However, the sequences of GPCRs contain some conserved residues and motifs. The well-established Ballesteros-Weinstein nomenclature[46] represents this finding by assigning the index '50' to the most conserved residue in each helix. For instance, Asn$^{1.50}$, denotes the asparagine in helix 1, which is the most conserved residue of this helix.

Careful examination of multiple sequence alignments of the different GPCR sequences reveals three additional motifs that occur with high propensity and have also been studied with regard to their functional roles. First, the L(I, M, V, T)xxxD (N, E) motif, where the latter one is D$^{2.50}$ and x is a non-ionic amino acid residue, most frequently A, S, L, or F.[47] Second, the D(E, N)R(K, H)Y(W, F, H) motif containing R$^{3.50}$.[48] And third, the N(D, T, S, Q)PxxY(F, W, H) motif including P$^{7.50}$.[49] Based on these motifs researchers can check if their sequence alignments are reasonable, i. e., whether these motifs are mapped correctly, as other sequential regions are too diverse across the whole GPCR family to give information about correctness.

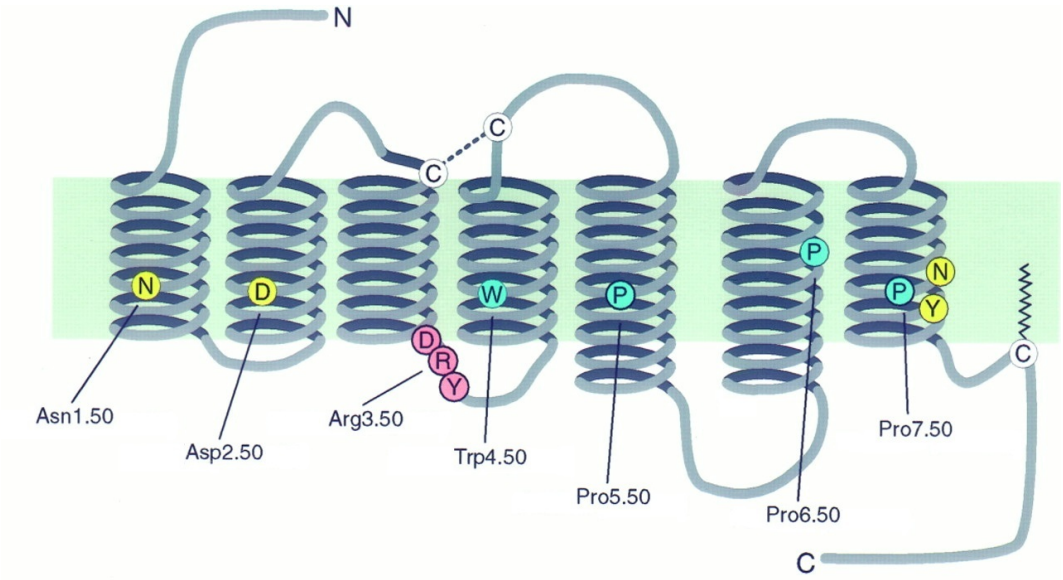


Figure 2.3: Two-dimensional model adopted after Gether.[50] The seven most conserved residues as well as the DRY and NPxxY motif are shown. D$^{2.50}$ belongs to the third motif LxxxD in helix 2. The almost invariable disulfide bond between C$^{3.25}$ and a cysteine located in the ECL2 is marked by a dashed line.

Moreover, a disulfide bond is found in almost all rhodopsin-like GPCRs, where one of the two cysteines involved is the conserved C$^{3.25}$ and the second is located

in the extracellular loop 2. The approximate positions of all mentioned residues and motifs are illustrated in Figure 2.3.

2.3 STRUCTURAL PROPERTIES

Several structural features are characteristic for the family of G-protein coupled receptors. First, the 7 transmembrane spanning helices are connected by loop regions, and are ordered counter clockwise when viewed from the extracellular side as illustrated in Figure 2.4. Helix 3 (H3) is the most tilted one, meaning that it is lying diagonally in the membrane, and has interactions with H5 at its cytoplasmic end. The N-terminus is always extracellular and the C-terminus intracellular. The latter domain is responsible for the activation of the bound G-protein. The ligand binding pocket is mainly located in the transmembrane region (H3-H7), however, the second extracellular loop (ECL2) plays a crucial role in ligand binding as shown in various studies, e. g., by Shi[51] or Massotte.[52] The inter-helical contacts are either highly conserved amino acids (polar, aromatic, or proline) or the small and/or weakly polar amino acids alanine, cysteine, glycine, serine, and threonine.[53] These interactions lead to a high packing formation of GPCRs. As mentioned above, another characteristic biochemical property is the conserved disulfid-bond between $Cys^{3.25}$ and a cysteine located in ECL2.

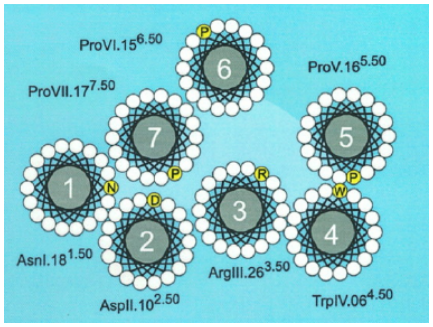


Figure 2.4: Counter clockwise ordered GPCR.[50]

2.4 CRYSTAL STRUCTURES

In this section, we will present the seven already solved (December 2011) crystal structures of G-protein coupled receptors and will describe briefly the most obvious differences between these structures with respect to the general structural features. Moreover, we will mention the functionality and the possible diseases each subfamily is associated with.

Bovine rhodopsin receptor

The first high-resolution (2.8Å) X-ray structure, bovine rhodopsin (PDB ID: 1F88[54]), was published in the year 2000. It reveals the major features of this protein family: the ECL2 (highlighted in yellow) forms a 2-stranded β-sheet and is extended into the binding pocket, where it forms several contacts with its covalently bound agonist retinal.[55] In subsequent years, the resolution has been improved to 2.2Å (PDB ID: 1U19[56]). Mutations of the rhodopsin gene are a major factor to various retinopathies, e.g. autosomal dominant retinitis pigmentosa[57] and night blindness.[58]

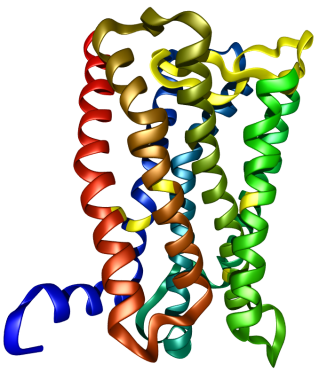


Figure 2.5: 1U19

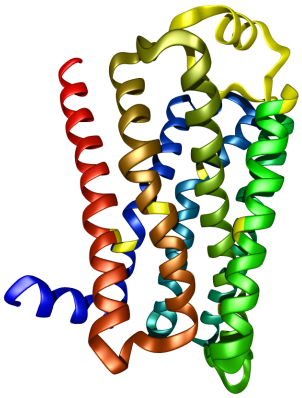


Figure 2.6: 2RH1

Human β_2-adrenergic receptor
Seven years after the first GPCR crystal structure was experimentally determined, the second one, a human β_2-adrenergic receptor (β_2AR), was resolved at a 2.4Å resolution (PDB ID: 2RH1[9]). The most surprising fact of this structure is a short helical segment in ECL2, which is stabilized by an additional intra-loop disulfid bond between Cys184$^{4.76}$ and Cys190$^{5.29}$, far above the binding domain. The bound antagonist carazolol[59] fills similar spaces in comparison to retinal in rhodopsin. Modifications of adrenergic receptor genes are associated with various diseases such as asthma, hypertension, and heart failures.[60]

Turkey β_1-adrenergic receptor
In 2008, the number of available GPCR structures was doubled. First, a turkey β_1-adrenergic receptor (β_1AR) was solved at a 2.7Å resolution (PDB ID: 2VT4[10]). This structure shows a distinctive kink in helix 1 (H1) of chain A, but the authors suggest chain B without a kink in H1 to be more reliable. Similar to 2RH1, a short helix is found in ECL2, from which researchers infer a common feature of βAR structures. In addition, a well-defined helix is observed in cytoplasmic loop 2 (CL2), which interacts with the highly conserved DRY motif at the end of H3. This might give researchers more insights into the activation process because former studies demonstrated the importance of CL2 and CL3 in G-protein activation.[61] Besides, the present ligand, the antagonist cyanopindolol, reveals the same binding mode as carazolol in β_2AR.

Figure 2.7: 2VT4

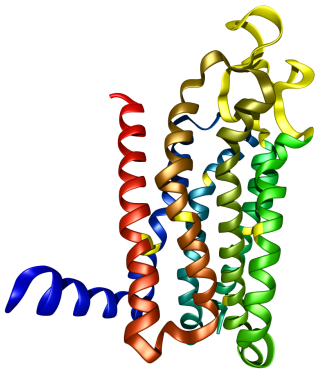


Figure 2.8: 3EML

Human A_{2A} adenosine receptor
The second structure resolved in 2008 was a human A_{2A} adenosine receptor (PDB ID: 3EML[11]). Remarkable differences are observed in the arrangement of the extracellular loops and in the binding of the antagonist ZM241385. The latter is due to a subtle replacement of the helices resulting in more ligand contacts with H6 and H7 and less interactions with H3 and H5. This finding disproves the assumption of a GPCR family specific binding pocket. The overall C_α-RMSD between this receptor and previously solved GPCRs lies between 2.0 to 2.5Å. Adenosine receptors are, among others, associated with pain regulations,[62] respiration,[63] and sleep.[64]

Human CXCR4 chemokine receptor

The evolutionary most diverse GPCR (see Figure 1.1) among the available crystal structures, a CXCR4 chemokine receptor (PDB ID: 3ODU[65]), was determined in 2010 and shows significant structural differences to the others. Shifts of the extracellular ends of H1, H4, and H6 as well as a 120° rotation of the extracellular end of H2 yield a different binding pocket formation. However, the C_α-RMSD to the other available structures is again between 2.0 to 2.2Å, while the extracellular half is more diverse ($\geqslant$2.2Å) than the intracellular one ($\leqslant$1.9Å). Chemokine receptors are regulating the migration of various cell types, e. g., leukocytes.[66–68]

Figure 2.9: 3ODU

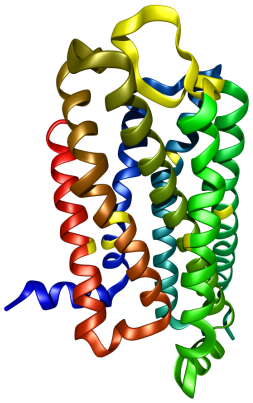


Figure 2.10: 3PBL

Human dopamine D3 receptor

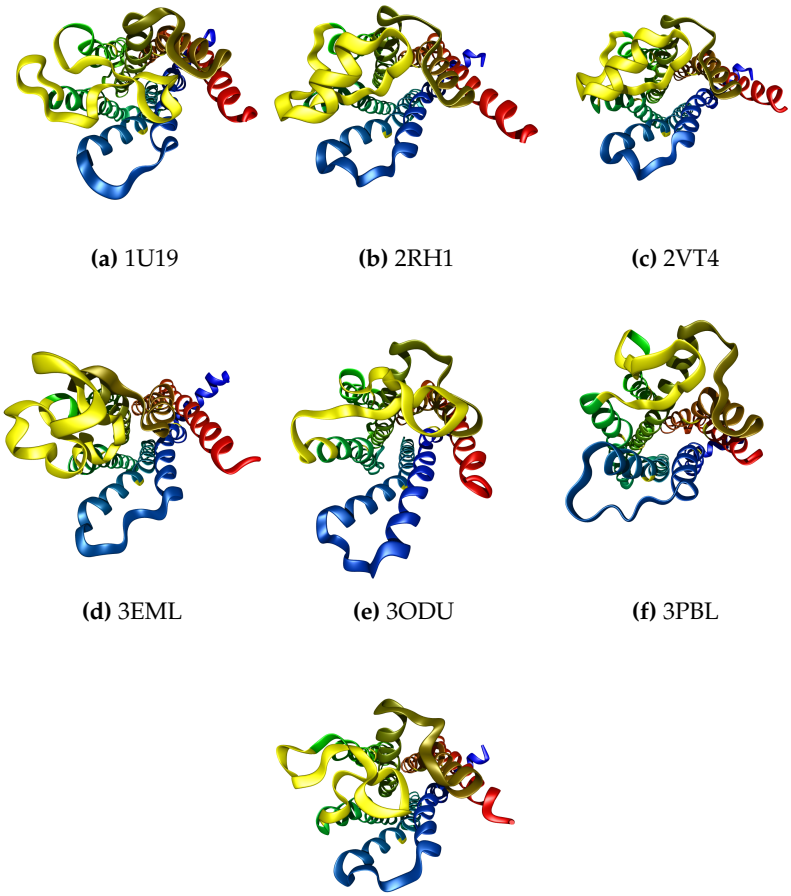
The structure of another GPCR, a human dopamine D3 receptor (PDB ID: 3PBL[69]), was published at the same time as 3ODU. The shorter ECL2 has no helical segment, but contributes to the ligand binding pocket in a similar way as βARs. Although a salt-bridge between $Arg^{3.50}$ and $Asp/Glu^{6.30}$ was assumed to be important in G-protein activation, this receptor is, besides rhodopsin, the only one, where this so-called *ionic lock* is present. Dysfunctions of this GPCR subfamily lead to several diseases in the central nervous system, e. g., Tourette's syndrome, schizophrenia,[22] and Parkinson's disease.[42]

Human histamine H_1 receptor

The last structure solved before writing this thesis was a human histamine H_1 receptor in 2011 (PDB ID: 3RZE[70]), which has a higher structural similarity to both aminergic receptors and the dopamine D3 receptor compared to the other three known crystal structures. A longer ECL2 section and an increased distance between the extracellular ends of H3 and H5 results in a larger volume of the ligand binding pocket, which is completely filled out by larger H_1R antagonists alleviating the symptoms of allergies and inflammation.[41,71]

Figure 2.11: 3RZE

Figure 2.12 shows the seven crystal structures from extracellular site. While the overall folds seem to be very similar (for better comparison see Figure 4.3), it is obvious that ECL2 is diverse across the different GPCRs. In 1U19 and 3RZE it covers the binding pocket to a large extent, whereas it is far above in the other five structures. Moreover, in case of adrenergic receptors (PDB ID: 2RH1 and 2VT4), it is even long enough to form a small α-helix, whereas it is very short in case of the dopamine receptor (PDB ID: 3PBL).



(a) 1U19 **(b)** 2RH1 **(c)** 2VT4

(d) 3EML **(e)** 3ODU **(f)** 3PBL

(g) 3RZE

Figure 2.12: The seven experimentally determined GPCR crystal structures as seen from the extracellular side. The ECL2 and the most conserved residue in each helix are marked in yellow.

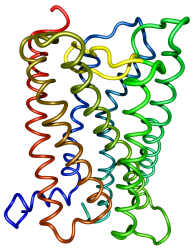
Table 2.2 gives an overview of all modifications that have been done to obtain more stable structures for crystallization. Note, none of the stabilizing procedures guarantees that the structures remain unchanged.

Table 2.2: Experimental modifications of the X-ray structures

Modification	1U19	2RH1	2VT4	3EML	3ODU	3PBL	3RZE
T4 lysozyme		X		X	X	X	X
Point mutation			X		X	X	
Modified termini		X	X	X	X	X	X
Inverse agonist	X	X					X
Antagonist			X	X	X	X	

The T4 lysozyme replaces most of the unstable third cytoplasmic loop. The mutations enhance thermal stability. Longer tails have just been deleted. The purification with ligands increases the stability, too.

3 *Homology Modeling*



Our first step towards automated G-protein coupled receptor modeling was the adaption of the commonly used homology modeling (HM) approach, which has successfully been applied to different globular[72] as well as membrane proteins.[73,74] The idea of HM, also known as comparative modeling, is based on the observation that three-dimensional structures of proteins are typically more conserved than their amino acid sequences.[75] Consequently, proteins with homologous sequences are expected to show a similar three-dimensional structure. In general, comparative modeling consists of four steps: template selection, sequence alignment, model building, and model validation.[76] The choice of an appropriate template structure is not always straight forward, in particular, if no template with a sequence identity of at least 30% to the protein of interest exists. In case of a missing template structure, *protein threading* is done, where a database of known structures for other proteins is queried to 'thread' the sequence through secondary structure elements like α-helices and β-sheets. Using TASSER, a well-established method for protein threading, all identified G-protein coupled receptors have been modeled.[29] While the C_α-RMSD between the model and the native structure of bovine rhodopsin for the core region (residues 32 to 323) is comparable to other studies (3.3Å), the authors achieved an excellent C_α-RMSD value of 2.1Å for the TM region.

In our study, published in the Journal of Medicinal Chemistry in 2009, we followed another approach and examined the benefit of the new template structure of the human β_2-adrenergic receptor (PDB ID: 2RH1) in homology modeling.[77] For years the only available structure was bovine rhodopsin (PDB ID: 1F88) and all homology models were based on this template structure. Whereas the extracellular loop 2 (ECL2) of 2RH1 has a complete different conformation compared to 1F88, the two TM folds are quite similar. In case this holds also for other GPCR, we should be able to create appropriate models without experimental knowledge and manual interaction during the modeling procedure. In contrast to other studies, which had to interact the modeling procedure manually, we used both available templates simultaneously, such that local structural similarities of both structures can be inferred to our unknown protein. The higher flexibility in the backbone, should lead to more diverse models such that at least one of these should fulfill all constraints based on experimental knowledge.

In a recent study, Nowak and co-workers extended the general framework of homology modeling by a molecular docking step to improve and facilitate the model building and validation process.[78] After generating a large amount of models by an automated procedure to sample the side chain conformational space, a potent ligand was docked to all models to identify those side chain conformations of the binding site that are advantageous for ligand binding. Afterward, the information of the docking runs was used to fine-tune a new set of models by restricting the determined residues to appropriate positions during the model building process. The ligand was docked in the new model set, and the ligand-receptor complexes were evaluated based on the CScore, which was used to finally determine a set of best fitting models, re-entering the docking procedure with 30 ligands. Despite some manual refinements following afterwards, the top-scoring ligand-receptor complexes already revealed the general binding motifs of the serotonin 5-HT_{1A} receptor. With this approach, Nowak et al. successfully modeled this aminergic receptor, yielding impressive results in terms of high enrichment factors in virtual screening approaches. Remarkably, no additional experimental information, e. g., from mutagenesis studies, has been used in the first steps of this procedure. In our opinion, this is currently one of the most promising methods for approaching automated modeling of G-protein coupled receptors.

Here, we apply Nowak's method to the human neurokinin-1 (NK1) receptor, a member of the neurokinin receptor family. The natural ligands of this family, the neurokinins, also termed as tachykinins, are small neuropeptides that are widely distributed within the peripheral and central nervous systems and are involved in neurotransmission and neuromodulation. Studies suggest that they are involved in various inflammatory and immune diseases.[79,80] Therefore, many antagonists targeting the NK receptors have been developed as therapeutic agents.[81]

Evers and Klebe already developed a homology model of the NK1 receptor based on the structure of bovine rhodopsin,[82] which was suitable to identify a novel submicromolar antagonist by virtual screening. Similar to the approach of Nowak, they started with a large number of initial models and used ligand information from a docking run to further improve the models. However, Evers and Klebe used more prior knowledge, e. g., about the conformation of the ligand, from the beginning. They assumed that the bioactive conformation is identical with its geometry in solid state and hence performed a rigid docking. In addition, they evaluated the docking complexes based on interactions derived from mutagenesis data rather than the corresponding docking score. Thus, the approach of Klebe and Evers requires strong manual interaction during the model building and refinement process, whereas we want to judge the feasibility of highly automated approaches to GPCR modeling.

Therefore, we explore if the procedure by Nowak et al. is generally transferable to nonaminergic GPCR modeling to yield suitable initial models for further refinement steps in reasonable time and with reasonable effort. Considering that this approach has originally been applied to an aminergic receptor (5-HT_{1A}), we expect significantly more intrinsic difficulties in our case because NK1 belongs to the

group of peptide binding receptors and the putative binding site of the endogenous ligand differs from the binding site of small molecule antagonists.

The second aim of our study is a deeper understanding of the influence of multiple different templates on the comparative modeling of GPCRs. Hence, we use the bovine rhodopsin structure and the recently resolved human β_2 adrenergic receptor structure as templates in the homology modeling step. Moreover, we combine both templates to extend the accessible conformational space, leading to larger backbone flexibility in the model building, not investigated in aforementioned studies. We assume that we can improve our models using multiple templates in an automated fashion. The final evaluation of the models is based on virtual screening techniques on a data set compiled from the literature as well as in house molecules.

3.1 METHODS

3.1.1 *Software*

All homology models presented in this work were generated using MODELLER 8v2, an established standard for comparative modeling.[25] MOE 2007.09 (Molecular Operating Environment) was used for the alignments as well as manually refinements of the models.[83] The protein-ligand docking was performed using Glide (Grid-based Ligand Docking with Energetics) in SP mode.[84] The chemical compound was drawn using Symyx Draw.[85] The alignments were formatted and represented using ALSCRIPT,[86] the graphics containing 3D structures were generated with BALLView,[87] the molecular viewer and modeling tool of the Biochemical Algorithms Library BALL,[88] version 1.2, and the enrichment plots were created with the statistical program tool R.[89]

3.1.2 *Alignments*

We computed the pairwise alignments of the human NK1 receptor sequence (UniProt ID P25103) first with bovine rhodopsin (PDB ID 1F88) and second with the human β_2 adrenergic receptor (PDB ID 2RH1, removing the lysozyme fusion protein) denoting the resulting alignments as R1 and B1, respectively. To study the impact of related sequences on the alignment, we also performed a multiple alignment of the human NK1 receptor with the human NK2 (UniProt ID P21452) and human NK3 (UniProt ID P29371) sequence. The result was then aligned with the bovine rhodopsin sequence (called R123). Repeating the procedure with the human β_2 adrenergic receptor did not change the results of the B1 alignment. Furthermore, we carried out a multiple alignment of the human NK1 sequence with both the human β_2 adrenergic receptor and bovine rhodopsin (called RB1). All alignments were computed in MOE using a gap start penalty of 7.0 and a gap extension cost of 1.0. Because of the low sequence similarity, we decided to use the BLOSUM 30 substitution matrix. We checked the plausibility of each alignment on both the mapping of conserved motifs and residues of class rhodopsin-like GPCRs as well

as the number of gaps appearing in helical regions, manually adjusting unfavorably aligned regions. The residues proposed to be involved in the binding mode are additionally denoted using the Ballesteros-Weinstein nomenclature.[46]

3.1.3 *Model Generation*

For each alignment, 300 homology models were generated by employing MODELLER, using the structure of bovine rhodopsin (1F88) and/or the human β_2 adrenergic receptor (2RH1) as templates. From the sequence alignment of the target protein and known template structures, MODELLER derives restraints expressed in terms of conditional probability functions (pdfs) for the target protein.[90] Optimizing the placement of the target protein coordinates in the molecular pdf with a conjugate gradient algorithm in combination with some nondeterministic steps, the program obtains slightly different models for the same alignment. Thus, the generation of a large number of models ensures a thorough sampling of the conformational space of the side chains of the receptor. Employing more than one template increases backbone flexibility and thus expanding the accessible conformational space.

3.1.4 *Docking*

The potent nonpeptide NK1 antagonist CP-96345[91] (see Figure 3.1) was flexibly docked into all generated models using Glide in SP mode. Glide first produces a rough initial guess to reduce the search space, followed by a torsion-angle optimization of the most promising initial candidates. The best results of this second stage are then refined by a Monte Carlo method to produce the predicted docking pose.[92] Finally, the best docking pose based on the Glide-score is chosen. Prior to docking, the ligand was optimized with the MMFF94 force field in MOE.

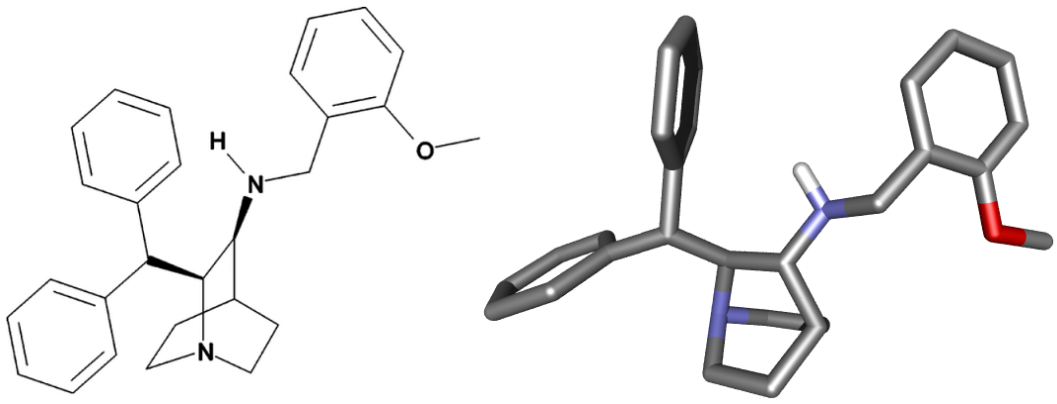


Figure 3.1: Structure of the quinuclidine amine **1** (CP-96345).[91]

3.1.5 *Model Refinements*

Following the approach of Nowak, we examined the top scoring docking poses to identify the essential key interactions that could be used to guide the model refinement. As shown in the Result and Discussion section 3.2 in detail, none of such interactions could be found prevalent in the top scoring docking poses. Thus, we had to visually inspect the docking results, taking further knowledge from mutagenesis studies into account. The most reliable suggestions concerning the binding mode propose an H-bond between the exocyclic secondary amine to Gln 165 (4.60) and an interaction between His 197 (5.39) and the benzhydryl group.[81,91,93–96] These findings from mutagenesis experiments are now taken as a substitute for the information, which was gained in the study by Nowak et al. by the first docking runs.
In a first refinement step, we restrained the χ_1 angle and the χ_2 angle of Gln 165 (4.60) to -60° and to 170°, respectively, to ensure its proper orientation into the binding pocket. The interaction between His 197 (5.39) and CP-96345 as well as the π-stacking between Tyr 272 (6.59) and His 197 (5.39) as suggested by mutagenesis experiments was strengthened by a clockwise rotation of helix 5 by 30° (seen from extracellular side) in the bovine rhodopsin template.
To further follow the approach by Nowak, we selected 14 conformationally diverse models from the restrained model set and docked a balanced set of 50 highly (IC50 < 1 μM) and weakly (IC50 > 10 μM) active NK1 ligands taken from the public database AurSCOPE to determine the most useful model for the identification of active ligands by docking. Unfortunately, but not unexpected, there was no model that separates the two groups satisfactorily. The failure of all models to separate the ligand groups can be attributed to the kind of interactions involved in the binding mode as well as the quite small activity difference for both ligand sets. Thus, the docking scores and also visual inspection did not give us additional information, which can be used for model improvement, and therefore we skipped further studies on multiple ligands, which is also in line with our aim to model GPCRs as automated as possible.
In a second refinement step manual changes, i.e., manual side chain placement of the binding site residues of three models followed and the selection process will be described in the next chapters. To relax the conformation of the generated docking poses, we performed an energy minimization of the side chain atoms employing the AMBER99 force field as implemented in MOE. The backbone atoms of the modified residues and the ligand atoms were kept fixed during this relaxation. Thereafter, we performed an energy minimization of the entire binding pocket and the docked ligand using the MMFF94 force field. All these refinement steps were done in MOE with default parameters.

3.1.6 *Virtual Screening*

For the virtual screening, we combined public domain ligands from the database AurSCOPE GPCR (company Aureus Pharma) and in house data of Boehringer Ingelheim to a set of 1784 molecules including 58 active ones. Active molecules are defined to have an IC50 value lower than 1 μM and all inactive ones have an IC50

value larger than 10 μM. The set was balanced among others with respect to the average molecular weight (actives: 463.7 Da; inactives: 425.5 Da) as well as average charge (0.414, 0.407) and the average number of rotatable bonds (6.88, 7.06) to minimize the influences of these parameters. The protonation states of the compounds were assigned using MOE, and the compounds were energy minimized with the MMFF94 force field before docking. The virtual screening was done with GLIDE in SP mode using default parameters and the enrichment plots were generated using R.

3.2 RESULTS AND DISCUSSION

3.2.1 *Alignment Study*

The overall sequence identity of the human NK1 receptor with bovine rhodopsin and with the human β_2 adrenergic receptor is lower than 30%. In this range, the number of alignment errors increases rapidly, resulting in the most substantial origin of errors in comparative modeling.[97] However, class I GPCRs share some highly conserved residues and motifs such that an unambiguous alignment can be achieved.[98–102]

Table 3.1: The four alignments used in this work

name	used sequences
R1	human NK1 and bovine rhodopsin
B1	human NK1 and human β_2 adrenergic receptor
R123	human NK1, human NK2, human NK3, and bovine rhodopsin
RB1	human NK1, bovine rhodopsin, and human β_2 adrenergic receptor

The names are composed of the used template structures bovine rhodopsin (R) and human β_2 adrenergic receptor (B) as well as the number of the human neurokinin receptor (1-3).

In all four alignments (see Figure 3.2 and Table 3.1), the conserved residues and motifs are correctly aligned, resulting in a proper arrangement of the seven helical regions. Moreover, 30 residues forming the general binding cavity for ligands identified by Rognan et al. in an extensive study of 369 nonolifactory human GPCR sequences are also correspondingly aligned in all four cases.[104] The alignment of the binding site residues of the human NK1 receptor proposed by the interaction model of Evers and Klebe, namely Gln 165, Glu 193, His 197, Ile 204, His 265, and Tyr 272, agrees with their published alignment,[82] except in the alignment R1. In this alignment, Tyr 272 (6.59) was mapped on Tyr 274 (6.57) of the bovine rhodopsin structure instead of the neighbored residue Phe 276 (6.59). This mapping was achieved by the insertion of a gap into the helical region. In our opinion, inserting gaps in structurally conserved regions is highly unlikely and indicates that a family alignment (as in the case of the multiple alignment R123) gives more reasonable results.

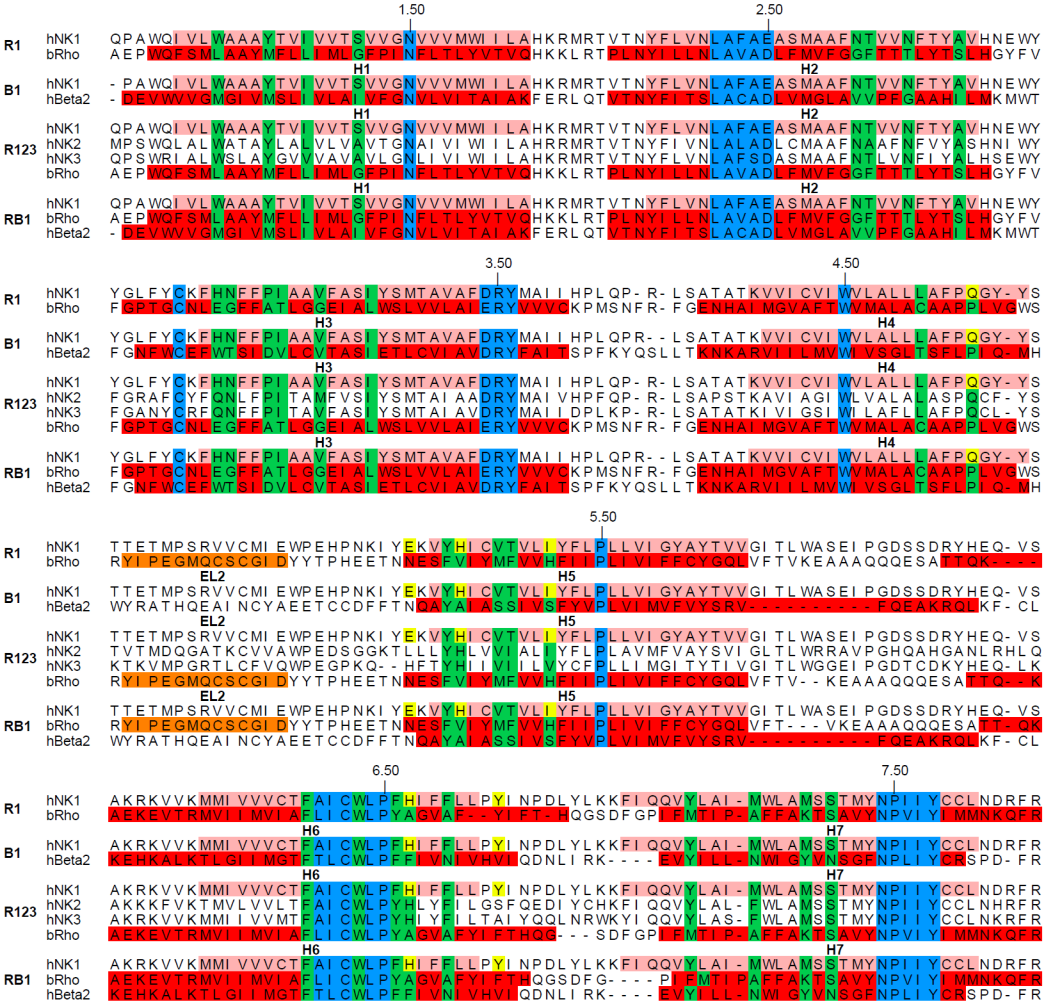


Figure 3.2: Sequence alignments used for model generation: First, the pairwise alignment of the human NK1 receptor and bovine rhodopsin (R1). Second, the pairwise alignment of the human NK1 receptor and the human β_2 adrenergic receptor (B1). Third, the multiple alignment of the human NK1-3 receptors and bovine rhodopsin (R123). Fourth, the multiple alignment of the human NK1 receptor, bovine rhodopsin, and the human β_2 adrenergic receptor (RB1). The red and orange marked regions are the TM helices of both templates and the ECL2 of the rhodopsin structure, respectively. The light-red marked regions indicate the TM helices of the human NK1 receptor predicted by TMpred.[103] Blue marked residues are conserved residues/motifs of class I GPCRs. Residues forming the TM cavity are marked in green.[104] All binding site residues of the human NK1 receptor, proposed by Evers and Klebe,[82] are colored in yellow. The alignments were formatted using ALSCRIPT.[86]

Table 3.2: NK1 residues involved in the binding mode

NK1	bovine rhodopsin (R123)	human β_2 adrenergic receptor (B1)	bovine rhodopsin (R1)
Gln 165 (4.60)	Pro 171 (4.60)	Pro 168 (4.60)	Pro 171 (4.60)
Glu 193 (5.35)	Asn 200 (5.35)	Asn 196 (5.35)	Asn 200 (5.35)
His 197 (5.39)	Val 204 (5.39)	Ala 200 (5.39)	Val 204 (5.39)
Ile 204 (5.46)	His 211 (5.46)	Ser 207 (5.46)	His 211 (5.46)
His 265 (6.52)	Ala 269 (6.52)	Phe 290 (6.52)	Ala 269 (6.52)
Tyr 272 (6.59)	Phe 276 (6.59)	Val 297 (6.59)	Tyr 274 (6.57)

The mapping of the NK1 residues involved in the binding mode of ligand CP-96345 and their corresponding amino acids based on the different alignments NK1bovine rhodopsin (R123) human β_2 adrenergic receptor (B1) bovine rhodopsin (R1).

3.2.2 *Structure Study*

To obtain reasonable orientations of the binding site residues in the homology model, these residues need to be mapped on residues of the template structure that are pointing into the binding pocket. As mentioned before the essential NK1 residues for binding of CP-96345 affirmed by various mutagenesis studies are Gln 165 (4.60) and His 197 (5.39).[93,95] Examining our alignments with bovine rhodopsin, these amino acids are mapped to Pro 171 (4.60) and Val 204 (5.39), respectively. Both residues are oriented into the binding pocket, and thus the alignment seems to be reasonable in this region. In the case of the human β_2 adrenergic receptor, the previously mentioned amino acids are mapped to Pro 168 (4.60) and to Ala 200 (5.39). While the latter is directed toward the binding pocket, the position of Pro 168 (4.60) does not seem to be suitable because it is oriented toward the neighbored helix 5. However, the positions of its neighbors Lys 167 (4.59) and Ile 169 (4.61) are even less appropriate. Altogether, we suppose that the orientation of helix 4 in rhodopsin seems to be a more suitable template than helix 4 of the human β_2 receptor. For helix 5, however, the opposite holds regarding the corresponding residues to His 197 and Ile 204 (see Table 3.2). Hence, we expected that a model generated by the combination of both templates and thus including backbone flexibility will perform best in the virtual screening experiment.
In the case of the pairwise alignment R1, the mapping of Tyr 272 in NK1 to Tyr 274 (6.57) in rhodopsin is less reasonable than the mapping to Phe 276 (6.59) as in the case of the multiple alignment because it is directed away from the TM cavity. Thus, we skipped the alignment R1 in the further modeling steps. Other residues being involved in the binding mode denoted by Evers and Klebe are Glu 193 (5.35), Ile 204 (5.46), and His 265 (6.52).[82] Table 3.2 lists the corresponding residues, which are all pointing well into the TM cavity (see Figure 3.3).
In addition, we examined the position of the ECL2 in the two template structures carefully because it is described in other studies that the ECL2 of rhodopsin causes many difficulties in the docking process.[82,105] Following the approach of Nowak,

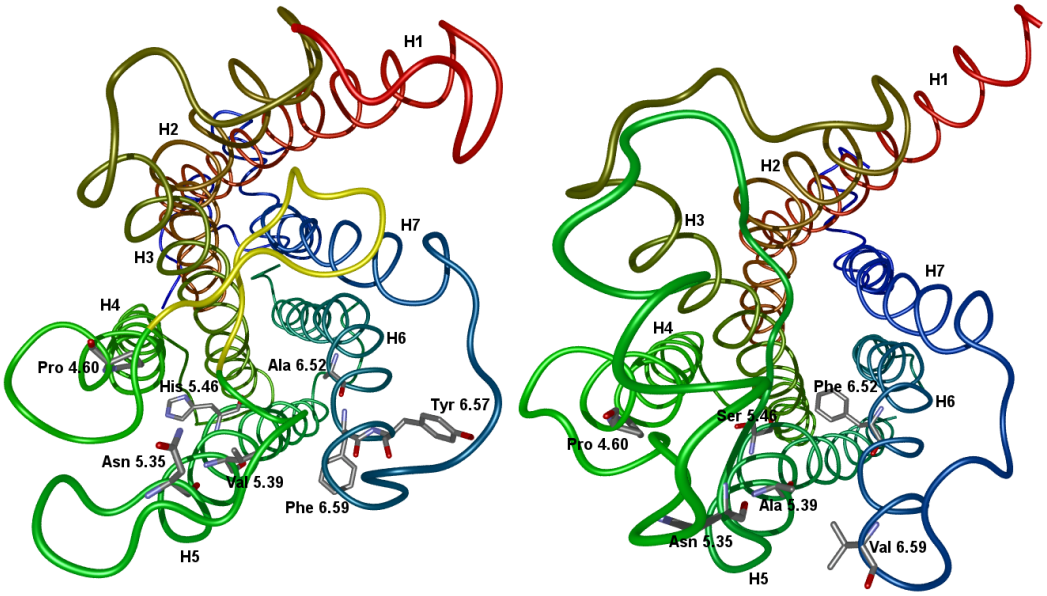


Figure 3.3: On the left side the template structure of bovine rhodopsin and on the right side the structure of the human β_2 receptor is represented. The removed ECL2 of the bovine rhodopsin structure is marked in yellow. All residues corresponding to the binding partners, proposed by Klebe, based on the alignments R123 (left) and B1 (right) are shown.

this extracellular loop was cut out of the homology model in a preprocessing step to ensure a successful protein-ligand docking. In contrast, the ECL2 of the human β_2 adrenergic receptor is located well above the TM cavity, forming a short helix rather than a β-hairpin. Hence, we did not expect significant difficulties caused by the ECL2 in the docking step. Consequently, we have cut out the ECL2 only in the homology models that are exclusively based on the template structure of bovine rhodopsin (alignment R123).

3.2.3 *Docking*

One of the best studied ligands for the NK1 receptor is **1**, and thus we used it for our first docking run into our initial models (Table 3.3, no. 1). Mutagenesis studies suggest that Gln 165 on helix 4 forms a hydrogen bond with the exocyclic secondary amine.[93] Furthermore, the binding affinity is negatively affected as soon as His 197 is mutated to alanine. The analysis of a series of **1** analogues identified the benzhydryl group as the binding partner.[94] Besides these two residues, various assumptions about other residues being involved in the binding mode, e. g., Glu 193, Ile 204, His 265, and Tyr 272, have been published.[81,82,95]
Because one of our aims was to test the general applicability of automated modeling procedures for nonaminergic GPCRs, we sorted the models according to their docking score of the best scoring pose as done by Nowak.[106] However, we could not identify essential key interactions in the docking complexes that could be used

to guide the model refinement by investigating the models and the corresponding score. Hence, solely from the docking pose and score, it is not possible to distinguish between reasonable and unreasonable homology models. We suppose that the reason for this result might be the different types of interactions. Nowak et al. modeled the serotonin 5-HT1A receptor, where strong interactions like salt bridges between the ligand and the receptor were formed during the docking procedure. In contrast, ligand binding in the human NK1-receptor involves only weaker interactions like hydrogen bonds or aromatic interactions. The scoring functions for docking do not seem to be sufficiently sensitive to properly rank these kinds of interactions, such that the score is not a good indicator for the quality of the docking poses in this case. Moreover, small changes in the conformation can yield large binding energy differences,[107] and because the scoring functions are adjusted based on the crystal structures, homology models perform in many cases worse than the corresponding crystal structures, especially if no strong interaction is involved in the binding mode. Therefore, the poses have to be inspected visually using additional experimental information, e. g., from mutagenesis studies as described in the next section.

Table 3.3: All model types generated in the modeling procedure

no.	name	alignment	template	restraints/refinements
1	INIT	R123	1F88	cut out ECL2
2	REST	R123	1F88	cut out ECL2 Gln 165: χ_1 = -60°, χ_2 = 170°
3	ROTA	R123	1F88	cut out ECL2 rotated helix 5 by 30° clockwise Gln 165: χ_1 = -60°, χ_2 = 170°
4	BETA	B1	2RH1	none
5	BOTH	RB1	1F88 + 2RH1	none
6	MO_INIT	R123	1F88	cut out ECL2 manually optimized binding pocket
7	MO_REST	R123	1F88	cut out ECL2 Gln 165: χ_1 = -60°, χ_2 angle = 170° manually optimized binding pocket
8	MO_ROTA	R123	1F88 rotated H5 30° clockw.	cut out ECL2 Gln 165: χ_1 = -60°, χ_2 angle = 170° manually optimized binding pocket
9	CONS	R123	1F88 rotated H5 30° clockw.	majority vote of 14 selected models
10	DEST	R123	1F88	cut out ECL2 manually destroyed binding pocket
11	RAND	R123	1F88	cut out ECL2

3.2.4 *Model Refinements*

Because the automated docking runs failed to identify crucial interactions between **1** and the receptor, we were forced to include mutagenesis data in the following refinement steps. We focused on the key interaction for binding CP-96345, a hydrogen bond between Gln 165 (4.60) and the exocyclic secondary amine of CP-96345 as well as an aromatic interaction between His 197 (5.39) and the benzhydryl group of CP-96345.

The first step was to reject all models, which do not agree with the mutagenesis studies. To this end, we used two simple filtering criteria: a distance filter between the Cδ of Gln 165 (4.60) and the exocyclic secondary amine of CP-96345 and a distance filter between the Cγ of His 197 (5.39) and the carbon atom of CP-96345 connecting the two benzene rings. Combined, these two filters reduced the overall number of models based solely on bovine rhodopsin (R123) to approximately 4% of all complexes. Although a distance filter of 5Åand 7.5Å, respectively, is very coarse, visual inspection confirmed that both residues point into the binding pocket and particularly to their postulated binding partners.

Closer inspection also showed that in the remaining models, the torsion angles of Gln 165 (4.60) have values of χ_1 = -60° (±5°) and β_2 = 170° (±5°). Thus, according to the original approach of Nowak, we constrained these angles for a new model generation run to optimize the distance to the binding partner of **1** (Table 3.3, no. 2). In the case of His 197 (5.39), we decided to modify the corresponding helix in the rhodopsin template by a clockwise rotation of 30°. This step was introduced to achieve both a strengthening of the interaction with **1** and to facilitate the formation of the π-stacking. Although this means a drastic change of the conformation, it has been shown by Vaidehi and co-workers that ligand induced changes of the backbone are quite usual.[108] We decided to rotate the helix directly in the template instead of postprocessing the generated models to avoid clashes in the postprocessing procedure. In these models, which are based on the modified template structure, we restrained the torsion angles of Gln 165 (4.60) as described above, too (Table 3.3, no. 3). This manual modification of the template, however, is not in the main focus of this study, which is the test of an automated GPCR modeling and virtual screening procedure. The modification of the template was done manually according to mutagenesis data, and therefore this technique may yield GPCR models closer to reality but is not suitable for an automated approach.

For the 300 models based on the human β_2 adrenergic receptor (B1, RB1), we were not able to identify common side chain features. In both cases, we used the abovementioned distance filters to reject those models, which do not agree with the mutagenesis studies. However, only a few of the models fulfill the filter criteria (<1%). In particular, for the models based on the alignment RB1, the backbones of the models vary too much such that a restraint or other modifications can not be applied to continue with a model refinement. For the subsequent virtual screening experiment, we selected from both model sets (Table 3.3, nos. 4 and 5) by visual inspection

a model among the top 10 scored complexes showing reasonable orientation of residues Gln 165 (4.60) and His 197 (5.39).

In a second refinement step, we selected from the pool of all generated models based on the R123 alignment 14 models manually. The selection was guided by the ability of various models to accommodate some public domain ligands as taken from the Aureus database. The models were picked according to reasonable docking poses and conformational diversity of amino acid residues close to the active site. To test the influence of the modeling techniques so far we selected three of these 14 models, one belonging to the INIT, one to the REST, and one to the ROTA model set (Table 3.3, nos. 6-8) to improve their interactions manually. This is not in line with an automated modeling process, but it will give us information about the general suitability of the various approaches for model generation so far.

Finally, we built a consensus model of all these 14 selected models (Table 3.3, no. 9). To this end, we took the backbone of a ROTA model and manually adjusted the conformation of the binding site residues to the conformation that occurs in most of all complexes, a procedure that is in line with an automated modeling process.

3.2.5 *Virtual Screening*

A virtual screening experiment was performed on a total of 11 different models (Table 3.3). Our two negative controls, an arbitrarily chosen model (RAND) from the INIT model set and a model with a 'destroyed' pocket (DEST) by manually manipulating the side chains such that they fill the binding site, worked as expected because no enrichment could be found (data not shown). Particularly, in the case of the model DEST (manually closed pocket), only a small amount was able to be docked into the binding pocket by GLIDE. Although more ligands (about two-third) could be docked into the model RAND, the enrichment factor is lower than random for this model.

Next, we compared the enrichment curves of the three nonoptimized models (nos. 1-3) with their corresponding manually optimized ones (nos. 6-8) (left picture in Figure 3.4). Although we expected an improvement, the manually optimized models yield a lower enrichment factor in two of the three cases. The reason for this effect depends on the optimization process itself and may be attributed to overfitting: since we improved these models manually to strengthen the contact between the important side chain residues and CP-96345, we simultaneously may have reduced the possible interactions to other ligands not taken into account. This led to a worse result in the virtual screening.

The analysis of the enrichment curves of the three other models (CONS, BETA, and BOTH) shows interesting results (right picture in Figure 3.4). In contrast to the manual refinements, the CONS model improved the results noticeable. Hence, the combination of the best side chain conformations is a reasonable step in model tuning. Its enrichment factor of the top 10% equals 2.6, which is in agreement with other virtual screening experiments of nonaminergic GPCR models.[109]

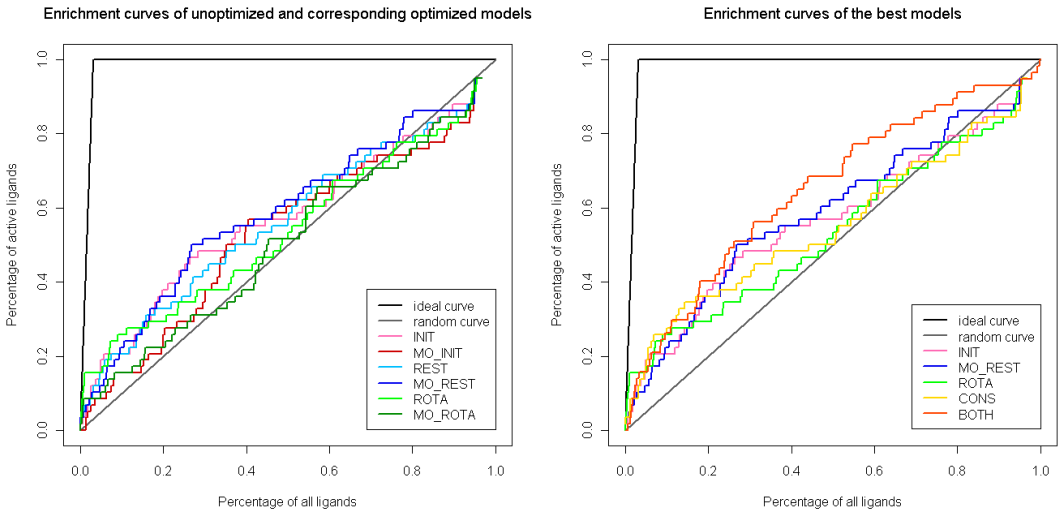


Figure 3.4: On the left the enrichment curves of the nonoptimized models and the corresponding manually optimized ones are shown. The figure on the right shows enrichment curves of the best models using the rhodopsin structure, the model using both templates, and the consensus model.

For the models based on β_2, we also expected an improvement for two reasons. First, the position of the ECL2 well above the transmembrane regions and, second, the appropriate orientation of helix 5. But the enrichment factor is surprisingly lower than random (curve not shown). We suppose that especially the position of Gln 165 on helix 4 might be crucial for the binding, but in this region, the β_2 adrenergic receptor is unsuitable as discussed in the section Structure Study 3.2.2. However, using the combination of the two templates β_2 and rhodopsin (BOTH), the enrichment curve equals the one of the consensus model and is in most cases even better (right picture in Figure 3.4). Remembering that this model was straightforwardly generated, e. g., without cutting out ECL2 or any refinement steps, this is a remarkable result. Most steps were performed automatically using scripting languages with the only (important) exception of the choice of the model out of the 300 generated. This manual effort was, however, negligible, because we have only looked at the models with the 10 best docking scores. Hence, in this case, the docking score guided us to find a reasonable model efficiently, although we were not able to discover important side chain conformations for further refinement steps as shown by Nowak et al.
In Figure 3.5, we represent the backbone of the model BOTH used in the virtual screening experiment. Here, the advantages of both templates are combined. First, the discussed positions of Gln 165 (4.60) and His 197 (5.39) are directed well into the binding pocket and, second, the previously mentioned π-stacking between Tyr 272 (6.59) and His 197 (5.39) is formed.

This result sheds light on the invaluable information the GPCR modeling community has gained by the resolution of the β_2 receptor and will gain by every newly emerging GPCR crystal structure.

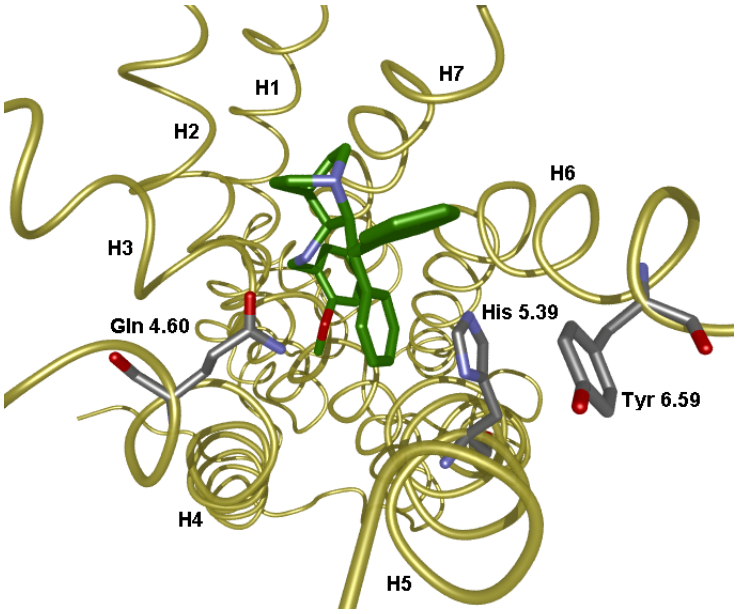


Figure 3.5: Backbone representation of the best model using two template. For a better view, we cut out the ECL2. The interactions with the two important residues Gln 165 (4.60) and His 197 (5.39) are accentuated.

3.3 CONCLUSION

In this study, we have investigated to which extent automated modeling of GPCRs is possible. Therefore, we followed the first modeling steps described by Nowak and co-workers. In our opinion, this is the most promising approach to improve homology models on an automated basis. We have thus attempted first to employ the approach to the modeling of the human NK1 based on the bovine rhodopsin template structure.

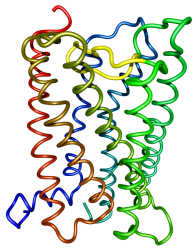
However, we soon found that this approach does not work for the nonaminergic case: our experiments found no essential key interactions in docking runs that could be used to guide the model refinement. We suggest that the reason for this insufficient result lies in the different type of interactions because in NK1 modeling no strong interactions such as salt bridges are involved in binding. In particular, aromatic interactions do not seem to be parametrized in an optimal way in current scoring functions and are therefore hard to identify. The high flexibility of CP-96345 complicated the docking procedure additionally. Hence, we had to include additional experimental information derived from mutagenesis studies from the very beginning. The refinements, especially the rotation of helix 5 in the bovine rhodopsin structure, improved the results significantly. This shows that for modeling GPCRs, we still rely on experimental data to generate promising models.

Employing the human β_2 adrenergic receptor as a single template, however, yielded unsatisfying results, such that even a manual refinement based on the docking results was not feasible. Nonetheless, the combination of both available

templates (rhodopsin and β_2) and the resulting expansion of the backbone conformational space for model building yields an enrichment factor in the range of the manually constructed consensus model. This was achieved without further refinements and just by choosing one of the top scoring docking complexes of CP-96345, whose side chain orientations are confirmed by experimental data. Thus, we suggest that the usage of multiple templates improves the models in a constitutive way. Hence, the human β_2 adrenergic receptor and probably also the other recently crystallized structures are very valuable for the homology model building of GPCRs. This shows that the availability of more structures improves the model building process because a larger conformational space, in particular in backbone regions, can be sampled. Using homology modeling in combination with docking, however, seems to be a viable option for automated receptor modeling if an essential very strong interaction between the ligand and the receptor is postulated, as in the case of amine receptors or fatty acid receptors.

Concluding our study, we suppose that data from mutagenesis studies must still be used to guide through the refinement steps of initial GPCR models. However, we suggest that these models should be generated based on multiple templates to include backbone flexibility. Hence, the crystallization of further GPCR structures has opened new ways to improve GPCR model building significantly.

4 *Ab initio Modeling*



In our previous study, we demonstrated that increasing backbone flexibility using a multiple template approach can replace manually defined restraints in homology modeling of G-protein coupled receptors (GPCRs). Our results are promising for automated in silico GPCR modeling, however, as mentioned before, only two crystal structures from the whole GPCR family were published, namely the bovine rhodopsin (PDB ID: 1F88)[54] and the human β_2-adrenergic receptor (PDB ID: 2RH1).[9] Although both structures have a similar fold in their transmembrane region (we will discuss this later in more detail), we could not infer that this holds also for other GPCRs, in particular, since both structures are evolutionary closely related as shown in the GPCR family tree (see Figure 1.1). If the folds are more diverse in other GPCRs, e. g., proteins of other GPCR subfamilies, our homology modeling approach might fail. Moreover, although we got reasonable results, there is still much potential for improvements.

To go further in the direction of automated GPCR modeling, we therefore focused on ab initio modeling approaches, where no template structure is needed. In contrast to homology modeling, all structural information has to be inferred from the target protein sequence.
In 2001, Shacham et al. developed PREDICT, which was particularly designed for GPCR modeling.[32] The authors demonstrated its apparent simplicity and success at the time by modeling rhodopsin and claimed to have modeled its binding pocket very accurately. From our experience in structure modeling, we supposed that the algorithm in its published state is hardly applicable for automated modeling of all GPCRs, but it should serve as a good basis for our own research. Hence, we had two goals: First, since the source code was not available (closed source), we had to re-implement the approach based on the information given in the corresponding publication and re-check its applicability to the rhodopsin structure.[33] Second, since we expected that PREDICT was going to fail in some instances, we aimed to replace the simple methods used, e. g., the three-dimensional brute-force optimization procedure, by more sophisticated ones to hopefully be able to build appropriate models of GPCRs without manual interference in the modeling process.

We first briefly describe PREDICT step by step and, when we go into detail for the most important parts, we present and discuss the results of our modified version of the PREDICT algorithm. To this end, we used all seven available crystal structures and manually re-checked the used prediction methods with respect to their general applicability in elaborate studies.

PREDICT has been tailored for GPCR modeling and thus, during its development several assumptions were deduced from the only structure that was available at the time (bovine rhodopsin). On the one hand, the TM helices have a length between 20 and 30 residues and are arranged in a counter-clockwise manner viewed from the extracellular (EC) side. On the other hand, these helices are connected by short loops such that their sequential order equals the order in 3D. Because these helices are embedded in a hydrophobic environment, Shacham and coworkers have forced the hydrophilic side chains to point into the interior part of the protein.

The PREDICT algorithm has been developed for arranging the helices only and did not focus on loop modeling. Therefore, the first step is to predict the seven TM helices from the sequence. The authors claimed that their algorithm does not depend on the determination of the exact helix location and that every prediction method, e. g., TMPRED,[103] PHDhtm[110] or TMHMM,[111] can be used. The extracted helices are projected to 2D along their axis and systematically arranged and optimized with regard to their orientation and inter-helical interactions. To save computational time, the helices build in 3D are then converted to a reduced representation defined by Herzyk et al.[36] In the following 3D optimization procedure including the vertical arrangement, the orientation, the helical center in the x-y plane, and the helical tilt angles, an adapted energy function is minimized using a simple brute force algorithm in a small range. After each optimization step, the side chains are optimized using a Monte Carlo algorithm and a rotamer library. The whole process stops when the energy is converged.

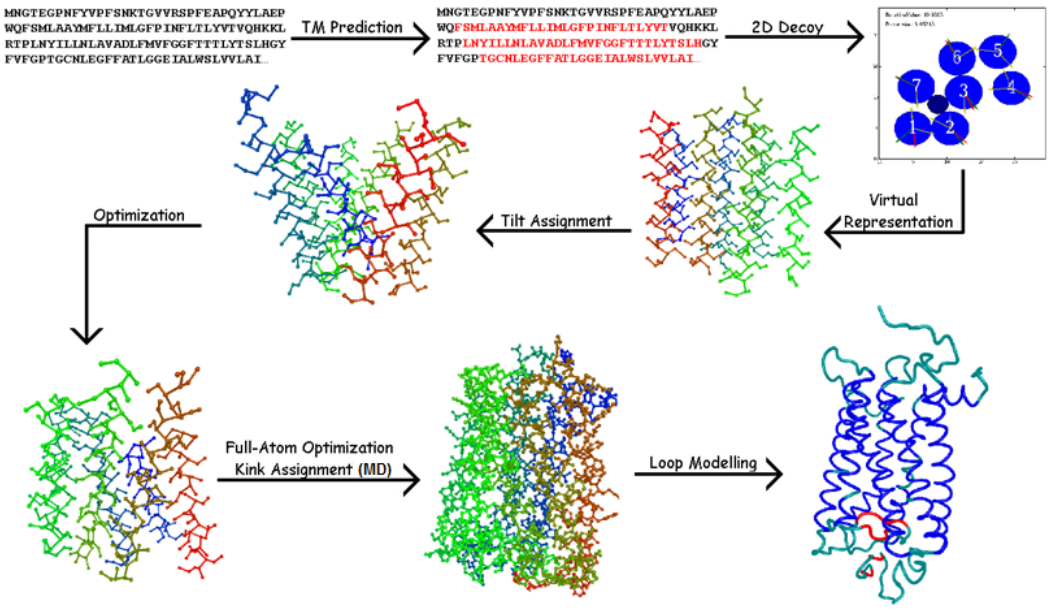


Figure 4.1: The PREDICT approach pipeline and subsequent loop modeling.

From the large number of decoys, Shacham et al. selected up to five models based on experimental data, converted them to a full atomistic model and minimized the structure based on the Consistent Force Field.[112] Finally, a molecular dynamics (MD) simulation is run to introduce kinks in each helix without changing the model

significantly. The pipeline of the whole PREDICT algorithm is sketched in Figure 4.1.

In the following sections, we will discuss four parts of the PREDICT algorithm in detail:

1. Secondary Structure Prediction (SSP)
2. Scoring function and optimization procedure in 2D
3. Scoring function and optimization procedure in 3D
4. Assigning kinks in helices

4.1 SECONDARY STRUCTURE PREDICTION

As stated above, the first step of PREDICT is the prediction of the seven transmembrane helices from the sequence. Becaus the helices exceed the TM region by far, most of them must be longer than 30 residues. This contradicts the first assumption of PREDICT, where helices are constrained to a maximal length of 30 residues. On the one hand, Shacham and coworkers suggested that PREDICT does not depend on the exact helix determination and we also agree that the length of the helices will not influence the algorithm significantly, but, on the other hand, side chains of residues in the first/last two turns of a helix in the EC side are involved in the binding of a ligand and might be missed by the SSP methods suggested by the authors, which focus mainly on determining the TM region of a protein. In our opinion, however, these residues are very important, not least in the validation of the final model by virtual screening. For this reason, we checked the performance of the three methods, PHDhtm, TMPRED and TMHMM with respect to the putative binding pocket residues.

A general GPCR binding cavity was derived from the retinal-bound crystal structure of rhodopsin (PDB ID: 1F88) in 2006 by Rognan and coworkers.[104] The authors identified 30 critical residues with a surface that is at least 25% accessible to a ligand (see Figure 4.2). The residues are annotated using the Ballesteros-Weinstein nomenclature and, hence, can be easily identified in other GPCRs. The findings described in the bachelor's thesis by Ernst concerning the applicability of this binding pocket definition to the other available GPCR crystal structures let us conclude that the extracted helices should be long enough to cover at least these residues.[113]

We applied all three SSP methods to the seven GPCR sequences and present the results for bovine rhodopsin (PDB ID: 1U19) and the human CXCR4 chemokine receptor (PDB ID: 3ODU) in Table 4.1; the complete results of all seven GPCRs can be found in the Appendix (Table A.1 to A.3).

Obviously, there are often (in 6 of the 14 illustrated cases) larger differences of more than one turn when comparing the prediction results of the three methods, e. g., the first residue of H7 of 1U19 (TMPRED: VII.27, PHDhtm: VII.34). This is very surprising, because Shacham and coworkers suggest that it is not important which

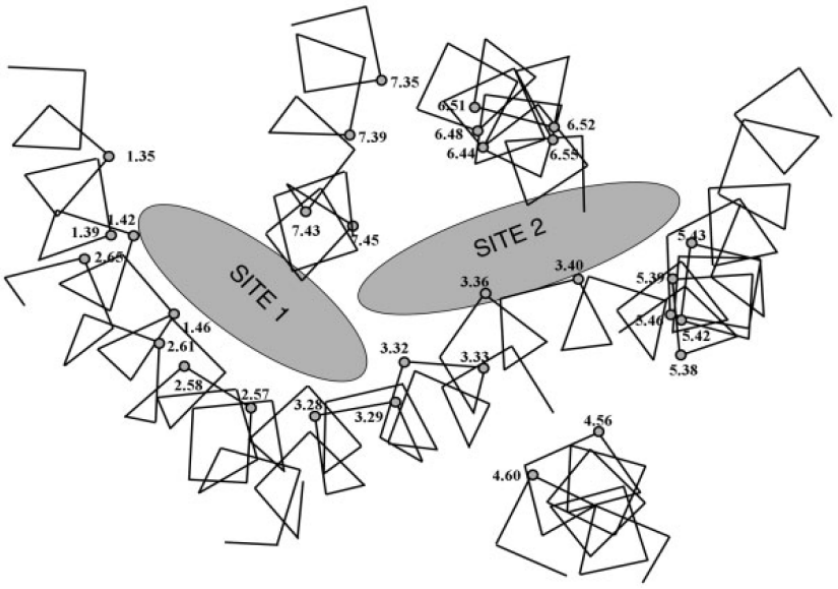


Figure 4.2: The critical points of a general binding pocket for GPCRs as shown by Rognan et al.[104]

SSP method is used. However, when changing the prediction method, a different GPCR region is modeled, which has to be kept in mind. Hence, the question arises, which of these methods is the best in terms of predicting helices from GPCR sequences, in particular with respect to the binding pocket residues. Using TMHMM, which predicts TM regions of only up to 23 residues in length, some putative binding pocket residues are missed, e. g., I.35 and II.65. This is, as already mentioned, because the helices in GPCRs exceed the TM region. Moreover, from the length of the helix, researchers can possibly infer information about its tilt angle. Therefore, it is advisable to use a method that does not predict a helix region of fixed length. TMPred misses two binding pocket residues, III.30 (in 1U19) and V.37 (in 3ODU). Although the length of an extracted helix is not fixed, the proportions are not predicted correctly. For example, H3, which is one of the longest helices in GPCRs, is predicted only with a length of 22 residues in case for 3ODU. When applying PHDhtm, the predicted length is even worse, since the determined helices are even too short (smaller than 20 residues) to cross the membrane in several cases.

How can we improve the prediction results to fix the problem of missing binding pocket residues? First, one can simply add one turn (3-4 residues) to the extracellular end of each helix, but this solution might fail for unknown GPCRs. Second, the application of methods optimized for secondary structure prediction, e. g., JPred,[114] might be more reliable. But again, this kind of prediction methods also requires a manual refinement in some cases (data not shown). Third, researchers often prefer consensus methods if the application of single ones fail in some instances. Unfortunately, this common strategy will not work for SSP as it misses, for example, the binding pocket residue V.38 in 3ODU, which is falsely classified in 2 of the 3 cases.

Table 4.1: Prediction results of SSP methods for 1U19 and 3ODU.

Pocket	Rognan	H1	H2	H3	H4	H5	H6	H7
	first residue	I.35	II.57	III.28	IV.56	V.38	VI.44	VII.35
	last residue	I.46	II.65	III.40	IV.60	V.46	VI.55	VII.45

	1U19	H1	H2	H3	H4	H5	H6	H7
TMHMM	first residue	I.34	II.41	III.26	IV.42	V.37	VI.37	VII.33
	last residue	I.56	II.63	III.48	IV.64	V.59	VI.59	VII.55
	helix length	23	23	23	23	23	23	23
TMPRED	first residue	I.32	II.41	III.30	IV.42	V.38	VI.36	VII.27
	last residue	I.58	II.66	III.55	IV.64	V.59	VI.57	VII.56
	helix length	27	26	26	23	22	22	30
PHDhtm	first residue	I.30	II.41	III.25	IV.42	V.38	VI.36	(VII.34)
	last residue	I.58	II.67	III.53	IV.64	V.64	(VI.55)	VII.56
	helix length	29	27	29	23	27	20	23

	3ODU	H1	H2	H3	H4	H5	H6	H7
TMHMM	first residue	I.37	II.44	III.27	IV.44	V.40	VI.36	VII.34
	last residue	I.59	II.62	III.48	IV.63	V.62	VI.55	VII.56
	helix length	23	19	22	20	23	20	23
TMPRED	first residue	I.33	II.44	III.27	IV.44	V.43	VI.36	VII.34
	last residue	I.57	II.71	III.48	IV.64	V.63	VI.57	VII.56
	helix length	25	28	22	21	21	22	23
PHDhtm	first residue	I.32	II.43	(III.28)	IV.44	V.37	VI.34	VII.38
	last residue	I.59	(II.64)	III.54	IV.62	V.64	VI.59	VII.55
	helix length	28	22	27	19	28	25	18

First, we give the first and last residue belonging to the binding pocket as defined by Rognan et al.[104] The results of TMPRED, TMHMM and PHDhtm are presented for 1U19 and 3ODU. In some cases, PHDhtm predicted one very long instead of two single helices. Here, we used the refinement method (PHDThtm) and set them in brackets. Binding pocket residues missed by a particular prediction method are marked red.

In summary, the SSP methods used in PREDICT are neither reliable nor satisfying such that we are forced to use a different strategy. Because the folds of the known GPCRs are quite similar, we examined if this is true for individual helices as well. From the results, we might be able to deduce general helix ranges in the seven GPCR crystal structures as it was done by Rognan and coworkers for their binding pocket. Figure 4.3 already illustrates the similarity of the folds, nevertheless, we take a closer look at the helical ends.

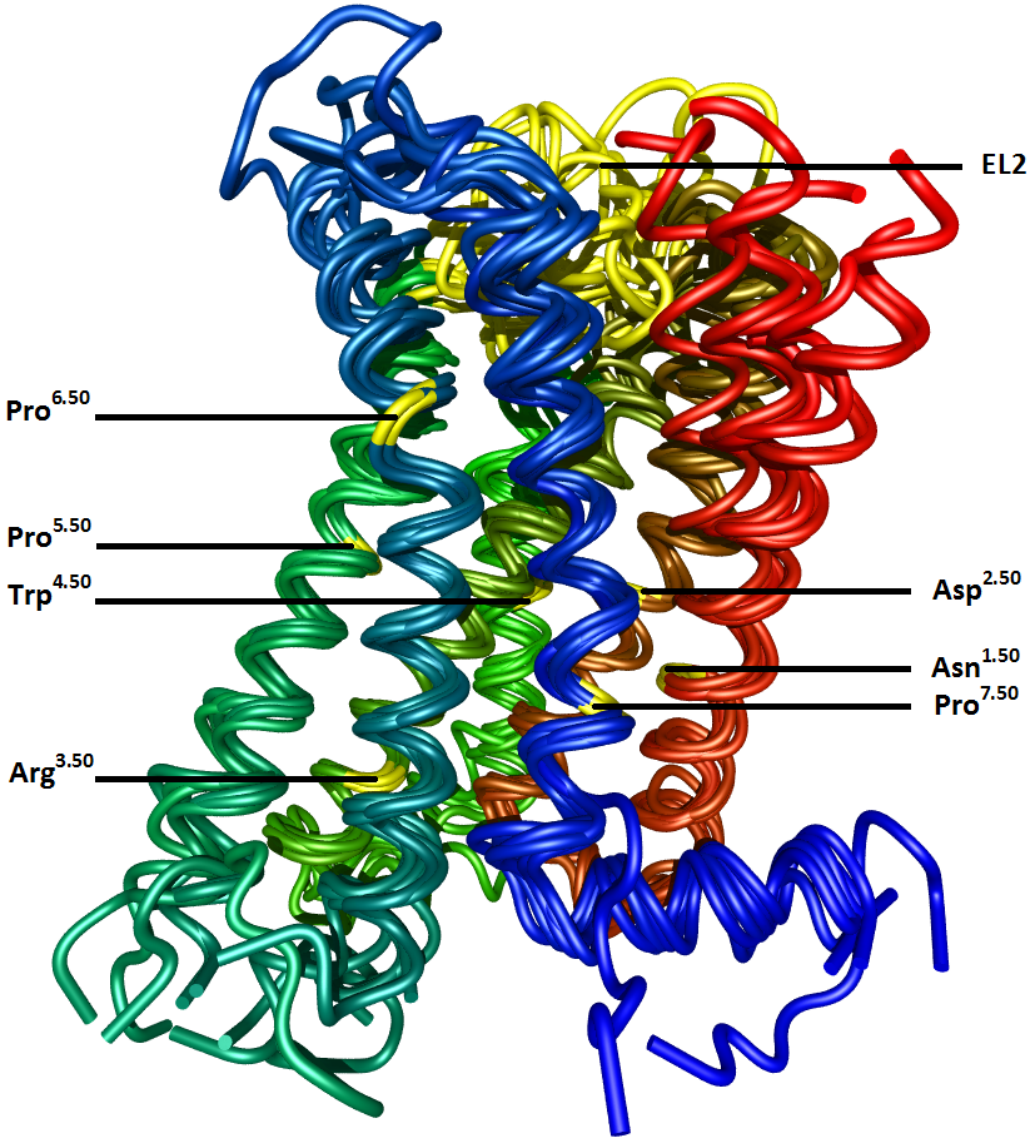


Figure 4.3: The GPCRs are mapped using the seven most conserved residues X.50 (marked yellow). The most diverse region is the extracellular loop 2 (ECL2), which is also highlighted.

To this end, we used BALLView to determine the first and last residue of each helix by visual inspection (see Table 4.2). Obviously, the lengths of the corresponding

helices in each structure are very similar. Those cases where the start and/or end of a helix differ by at least 3 residues are highlighted in red and will be discussed in comparison and also with regard to the ligand binding pocket located in the outer membrane side domain.

Table 4.2: Transmembrane helix regions identified by visual inspection

PBD ID		First	Cons.	Last		First	Cons.	Last
1U19	Helix 1	Glu33^{28}	Asn55	Gln64^{59}	Helix 2	Pro71^{38}	Asp83	His100^{67}
2RH1		Asp29^{28}	Asn51	Lys60^{59}		Val67^{38}	Asp79	Met96^{67}
2VT4		Trp40^{31}	Asn59	Ser68^{59}		Leu75^{38}	Asp87	Arg104^{67}
3EML		Ile3^{29}	Asn24	Asn34^{60}		Val40^{38}	Asp52	Ser67^{65}
3ODU		Ala34^{28}	Asn56	Gln66^{60}		Met72^{38}	Asp84	Val99^{65}
3PBL		Tyr32^{35}	Asn47	Glu57^{60}		Thr63^{38}	Asp75	Thr92^{67}
3RZE		Met28^{33}	Asn45	Glu55^{60}		Val61^{38}	Asp73	Leu89^{66}
1U19	Helix 3	Thr108^{23}	Arg135	Val139^{54}	Helix 4	Asn151^{40}	Trp161	Val173^{62}
2RH1		Phe104^{24}	Arg131	Thr136^{55}		Asn148^{40}	Trp158	Gln170^{62}
2VT4		Ser111^{22}	Arg139	Thr144^{55}		Arg155^{39}	Trp166	Met178^{62}
3EML		Cys74^{24}	Arg102	Ile108^{56}		Gly118^{39}	Trp129	Leu141^{62}
3ODU		Asn106^{22}	Arg134	Val139^{55}		Gln145^{34}	Trp161	Ile173^{62}
3PBL		Ile101^{23}	Arg128	Met134^{56}		Cys147^{39}	Trp158	Phe170^{62}
3RZE		Pro98^{23}	Arg125	Gln131^{56}		Thr140^{38}	Trp152	Gly164^{62}
1U19	Helix 5	Glu201^{36}	Pro215	Thr229^{64}	Helix 6	Thr243^{26}	Pro267	His278^{61}
2RH1		Gln197^{36}	Pro211	Gln229^{68}		Lys267^{29}	Pro288	Gln299^{61}
2VT4		Arg205^{39}	Pro219	Gln237^{68}		Arg284^{29}	Pro305	Asn316^{61}
3EML		Asn175^{36}	Pro189	Arg205^{62}		Ser210^{25}	Pro235	Cys246^{61}
3ODU		Leu194^{33}	Pro211	Lys225^{64}		His232^{28}	Pro254	Leu266^{61}
3PBL		Pro186^{36}	Pro200	Lys216^{66}		Leu322^{28}	Pro344	Cys355^{61}
3RZE		Thr188^{36}	Pro202	Val217^{65}		Asn225^{28}	Pro247	Phe257^{60}
1U19	Helix 7	Ile286^{33}	Pro303	Met309^{56}				
2RH1		Lys305^{32}	Pro323	Ser329^{56}				
2VT4		Asp322^{32}	Pro340	Ser346^{56}				
3EML		Trp255^{33}	Pro272	Arg278^{56}				
3ODU		Gln275^{26}	Pro299	Ala303^{53}				
3PBL		Pro362^{32}	Pro380	Phe386^{56}				
3RZE		Glu264^{32}	Pro282	Cys288^{56}				

The table gives the residue name in 3-letter-code as well as the residue ID written in the PDB file. The small numbers give the corresponding Ballesteros-Weinstein identifiers. Those differing by more than two residues are marked in red.

Helix 1 (H1) is well defined in all crystal structures, but the starting residue is in three cases up to two turns later. The main reason is that this helix is not completely contained in the corresponding PDB files. It is assumed that H1 has a general size of 33 residues starting from I.28 to I.60, but since the first residues are far away from the binding pocket cavity - and also missing in some structures - it is sufficient to

use a general region from I.31 to I.60 to model H1. H2 and H3 differ by no more than two residues and hence, have a common range from II.38 to II.67 and from III.22 to III.56, respectively. In both cases, we took the longest range. The only exception in H4 is the starting point of 3ODU. The helical structure shows a kink towards H3 after the first three turns, resulting in a very short loop between H3 and H4 as compared to the other structures. Since this part is located in the intracellular (IC) side, it is not critical for the modeling of the binding pocket. Hence, we defined H4 from IV.39 to IV.62. The EL2 is the longest and most diverse loop region in GPCRs and located between H4 and H5. While the starting point in all structures is IV.63, the end point has two exceptions. When taking V.36 as a starting position for H5 we take three additional residues in 2VT4 into account, while we shorten H5 in 3ODU. If taking into account that these residues point mainly toward the membrane, we can neglect them when defining a general helix region without introducing an error. The end position of H5 and the starting position of H6 are very different in each structure because a T4 lysozyme (T4L) fusion was inserted between these helices at the cytoplasmic side of the receptor to increase the stability during the crystallization process. Again, exact modeling of this region is not critical such that we defined the general helical range of H5 and H6 from V.36 to V.66 and from VI.28 to VI.61, respectively. Finally, H7 is defined from VII.32 to VII.56, where only 3ODU has a much longer helix compared to the others. Since these residues are far above a supposed binding pocket, we can neglect them for the purposes of our analysis. The common helix regions we infer from this discussion are presented in Table 4.3.

Table 4.3: The common helix region in GPCRs

	Manual	H1	H2	H3	H4	H5	H6	H7
Helix	first residue	I.31	II.38	III.22	IV.39	V.36	VI.28	VII.32
	last residue	I.60	II.67	III.56	IV.62	V.66	VI.61	VII.56
	helix length	30	30	35	24	31	34	25

The helices annotated manually by visual inspection. H4 and H7 are the shortest with a length of 24 and 25 residues, respectively. The other helices are composed of at least 30 residues with H3 being the longest.

A closer look at the positions of the binding pocket residues in the annotated helices shows that they are very close to the extracellular ends. H2, H4 and H5 end/start only two residues away from a binding pocket residue. When using the three suggested prediction methods, we should therefore not use a consensus method but set all residues to a helix if it is at least once defined as helix. Hence, we let predicted helix 'states' dominate non-helix 'states'. For the cases we discuss here, none of the putative binding pocket will then be missed. Nevertheless, the so predicted helices for bovine rhodopsin and the human CXCR4 chemokine receptor (see Table 4.4) are still shorter than our manually annotated ones. In future work on GPCRs of another subfamily, it is thus still much more likely to miss an important residue when using these prediction methods than our defined common helix regions.

Table 4.4: Helix regions identified by TM prediction methods

	SSP methods	H1	H2	H3	H4	H5	H6	H7
1U19	first residue	I.30	II.41	III.25	IV.42	V.37	VI.36	VII.27
	last residue	I.58	II.67	III.55	IV.64	V.64	VI.59	VII.56
	helix length	29	27	31	23	28	34	30

	SSP methods	H1	H2	H3	H4	H5	H6	H7
3ODU	first residue	I.32	II.43	III.27	IV.44	V.37	VI.36	VII.34
	last residue	I.59	II.71	III.54	IV.64	V.64	VI.59	VII.56
	helix length	28	29	28	21	28	34	23

The helices predicted using the TM prediction methods as obtained when considering a residue to be helical if at least one prediction method defined the residue as helix.

4.1.1 *Conclusion*

To summarize our findings, we strongly recommend not to use a single TM prediction method for defining the helix regions in GPCRs as suggested by Shacham and coworkers, for two reasons. First, the prediction methods often vary slightly but in a few cases even more than one helical turn. Applying different SSP methods leads therefore to different models that are not exactly comparable to each other. Moreover, due to the close location of binding pocket residues to the extracellular helical ends, there are a few examples where these residues are missed. Hence, the results of subsequent model validation methods, e. g., by virtual screening, are not reliable anymore.
There are two possibilities to handle the previously mentioned difficulties. On the one hand, the user can apply all three methods and take the longest range for each helix. At least in the cases we tested, none of the binding pocket residue is missed. On the other hand, a manual comparison of the available crystal structures, in particular, with regard to the helices, shows that a general helix region can be inferred. This helix region is slightly longer compared to the result when applying all three methods and is hence, not so prone to misclassification in unknown GPCRs. In addition, the models can easily be compared in this core region due to having the same length. To use this general helix region is consistent to the strategy of Rognan et al., who defined a general binding pocket cavity for GPCRs.

4.2 SCORING FUNCTION AND OPTIMIZATION PROCEDURE IN 2D

After extracting the helices, we need a first guess of their arrangement. For this task, Shacham et al. generated a large amount of two-dimensional decoys fulfilling some constraints, e.g., a maximal diameter of the molecule. The authors also applied some filters based on mutagenesis data to reduce the number of decoys. Figure 4.4 illustrates two examples of such a decoy. The seven helices are arranged in a counter-clockwise manner when viewed from the EC side. The distance between H2 and H7 is at most 20Å due to a conserved interaction between two residues of these helices.[115] We added another constraint forcing the center of H3 to be inside the convex hull of the other helices. In this experiment, approximately 200 two-dimensional conformations have been generated, but this number depends strongly on the mesh size of the grid. We then generated canonical helices from the

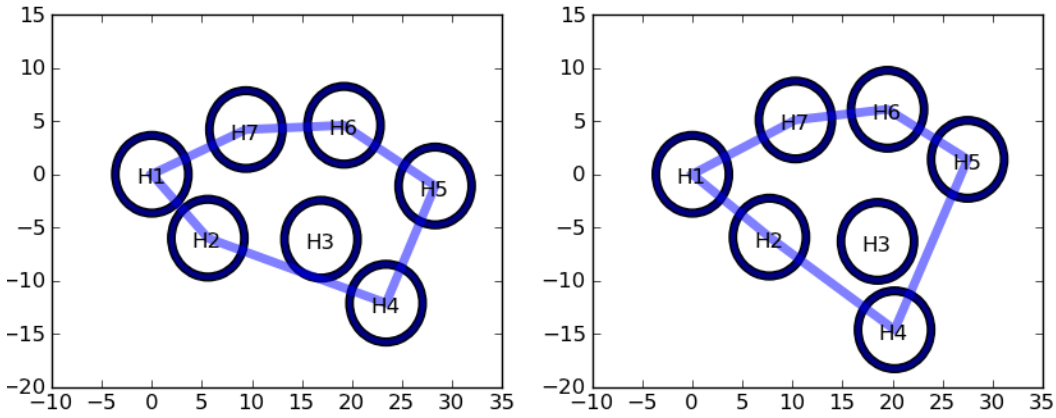
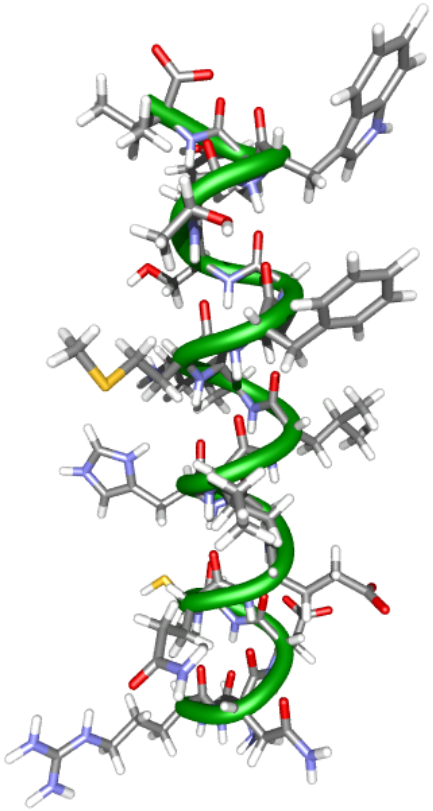


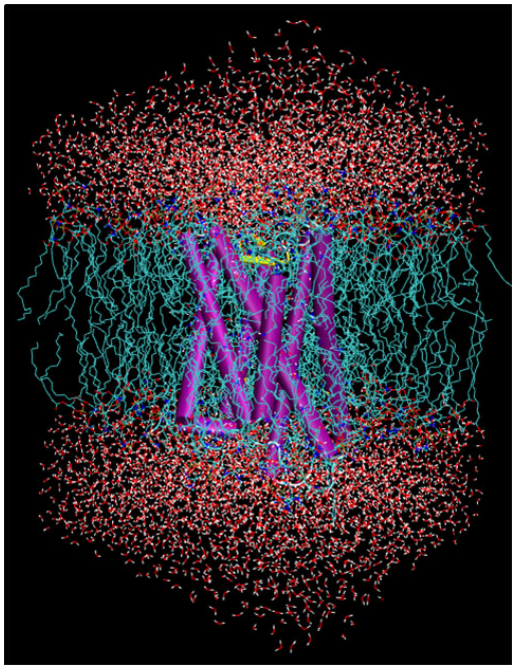
Figure 4.4: Two examples of 2D decoys. The helices are ordered counter-clockwisely when viewed from the EC side. H2 and H7 are close in distance. H3 does not belong to the convex hull (light blue polygon) of the helical centers.

given sequences using the *PeptideBuilder* class in BALL. Canonical helices feature consistent torsion angles of $-57°(\phi)$ and $-47°(\psi)$, such that 3.6 residues on average form one turn. The distance in axis-direction of two residues is 1.5Å, such that at least 20 residues are needed to cross a membrane. A canonical helix consisting of all 20 different amino acids is represented in Figure 4.5a.

Following the PREDICT approach, we used the hydrophobicity of amino acids as well as aromatic-aromatic helical interactions to find a first orientation of the helix bundle. In general, hydrophilic side chains of membrane proteins tend to be located at the inside, i.e., facing towards the other helices, while hydrophobic side chains point towards the membrane. However, since the membrane is a bilayer, this behavior is inverted in the phospholipid head group domain. Thus, using a simple summation of hydrophobicity values along those parts of the helix facing outward would be too simple as a score. For a reliable computation of the hydrophobic moment of a helix, we need information about the exact location in the membrane (as illustrated in Figure 4.5b) to determine whether a given residue would face a hy-



(a) Canonical α-helix



(b) GPCR embedded in a membrane

Figure 4.5: (a) Canonical α-helix consisting of all 20 different amino acids. (b) A GPCR embedded in a membrane. Image courtesy of Dr. Alpeshkumar Malde, School of Chemistry and Molecular Biology, University of Queensland.

drophilic or a hydrophobic part of the membrane, but this information is not available at the current state. Shacham and coworkers suggested to use a trapezoidal weighting mask, such that the hydrophobic moment $\vec{\mu}$ of a helix is dominated by residues located in the lipid part of the membrane and is computed as follows:

$$\vec{\mu} = \frac{1}{m} \sum_{i=1}^{m} w_i H_i \vec{S_i} \tag{1}$$

where m is the number of residues, w_i the weight, H_i the hydrophobicity scale and $\vec{S_i}$ the unit vector from the axis to the C_α-atom of the i-th residue. For the second parameter of Eq. 1, the hydrophobicity scale, we can choose between several different scales that have been proposed in the literature, among others, the one of Kyte and Doolittle or Eisenberg.[116,117]

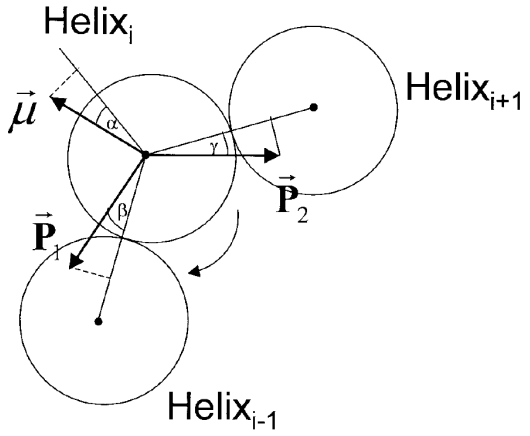


Figure 4.6: Schematical illustration of the three vectors $\vec{\mu}$, $\vec{P_1}$ and $\vec{P_2}$ used to optimize the initial orientation of helix i. This figure is taken from the publication of Shacham et al.[33]

The second term of the 2D scoring function includes the aromatic inter-helical interactions. Similar to Eq. 1, two vectors for these kind of interactions are computed as follows:

$$\vec{P_x} = \frac{1}{N_x} \sum_{i=1}^{n_x} \omega_i w_i \vec{S_i} \quad x = 1, 2 \tag{2}$$

where

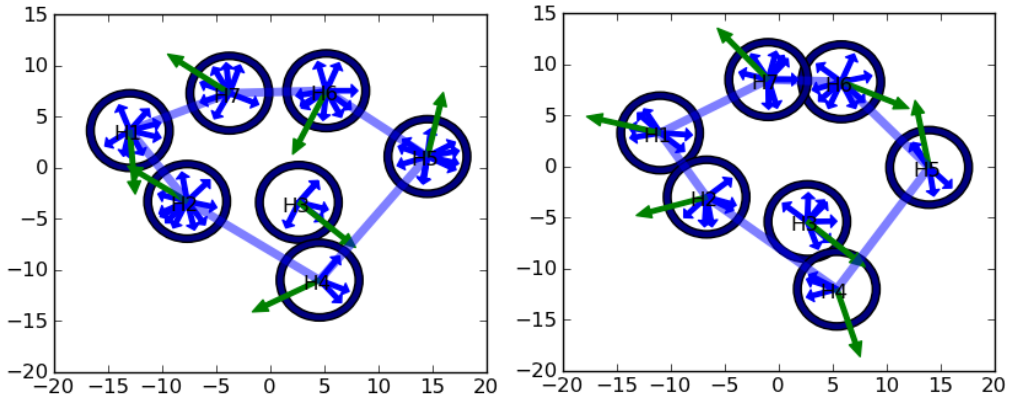
$$\omega_i = \begin{cases} 1, & \text{residue } i = \text{Phe, Trp, Tyr, His} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and n_x for $x = 1, 2$ is the number of residues facing either helix $h-1$ or helix $h+1$. To find a suitable orientation for each helix, Shacham et al. optimized

$$S = w_\mu \vec{\mu} \cos\alpha + w_{p1} \vec{P_1} \cos\beta + w_{p2} \vec{P_2} \cos\gamma \tag{4}$$

where w_μ, w_{p1}, and w_{p2} are the weights for the corresponding vectors and α, β, and γ the angles between the current vectors and the ideal one. The orientation optimization is done in brute force fashion, using 2° increments, and is performed separately for each helix to yield a greedy heuristic optimization procedure. Although this procedure seems to be straightforward and simple, the results are crucial for the following 3D optimization process of PREDICT since the 3D optimization method is based on a brute force algorithm and searches for an optimal orientation of a helix only in a small range of ±15°. Hence, a wrong initial orientation of more than 15° away from the correct solution will never lead to a correct conformation in the 3D optimization process.

Since Shacham et al. give no information about the values for the parameters w_μ, w_{p1}, and w_{p2} in Eq. 4, we had to fit these values ourselves. To obtain a first impression of the optimal starting orientation and to possibly infer suitable values



(a) 1U19 **(b)** 3ODU

Figure 4.7: The canonical variant of 1U19 and 3ODU in 2D. The helices in the crystal structure were replaced by their canonical version. The x,y-center is the average of the backbone atoms. The hydrophobic moment (green arrow) is computed using the hydrophobicity scale of Eisenberg and trapezoidal weighting factors. All inter-helical vectors are drawn as blue arrows.

for these parameters, we mapped each of the canonical helices to their corresponding in the crystal structure and computed their hydrophobic moment and all inter-helical vectors as specified in Eq. 1 and Eq. 2. Figure 4.7 illustrates the results using the PREDICT parameters (hydrophobicity scale of Eisenberg and trapezoidal weighting factors) for two GPCR structures.
Unfortunately, and very surprisingly, the hydrophobic moments do not always point towards the membrane in the crystal structures. For example, the corresponding vectors of H5 and H6 face each other in 3ODU. This observation can also be found in most of the other crystal structures and in a few cases between H1 and H2 (see Table A.1). Even more problematic is the fact that the hydrophobic moment of H6 in 1U19 points to the interior of the protein.
Thus, since these results indicated that the original approach will not lead to suitable starting conformations, we tested also other hydrophobicity scales found in the literature[118–122] and adapted the trapezoidal weighting factor using different Gaussian functions to reduce the impact of residues at the helical ends in a slightly different way. However, the best results we obtained include the hydrophobicity scale of Eisenberg, which is in agreement with PREDICT, while the different Gaussian functions did not influence the results significantly. Finding a meaningful setup for H6 in 1U19 corresponds to a worse result for H2 in 1U19. For different helices, different parameters would be optimal, however not necessarily across all known GPCRs. None of the parameter sets we tested could guarantee suitable solutions for all known GPCRs, let alone the unknown structures. Furthermore, the aromatic residues do not help, because they point in many diverse directions. Hence, when optimizing the 2D scoring function of PREDICT, we would be too far away from the correct orientation - independent of the choice for the weights in Eq. 4. Although we

think that this simple method has some potential, for example it works almost perfectly for 3ODU for all helices except H5 and H6, the decisive point is that we were not able to produce reliable results that we can use for the following optimization process - not even for rhodopsin, which was used as the example in the PREDICT publication.
The reasons for these results might be diverse. On the one hand, the hydrophobic moment is computed for canonical helices, but the ones in the crystal structure are often kinked, e. g., H6. This distortion is not considered in this step. On the other hand, although the impact of residues at the helical ends is reduced, they might still insert some noise in the calculation. The exact position in the membrane bilayer (including kinks and tilt angles) is also not known in the current modeling step. Moreover, the hydrophobic moment is the sum of the vectors from the helix axis to the C_α atoms, which might differ from the orientation of the corresponding side chain.

4.2.1 *Conclusion*

Finding a good start conformation, in particular an adequate helix orientation, for the following 3D optimization procedure is a crucial step in the PREDICT algorithm. The method to predict the orientation described by Shacham et al. depends on the hydrophobic moment and putative inter-helical interactions. Although some parameters used in the original approach are not available, and thus the approach denoted in the publication of PREDICT is not entirely reproducible, we doubt that their simple method is suitable for all GPCRs. As shown above, the optimal starting orientation does not fit to the 2D energy function, independent of the choices of the missing parameter values. It is thus highly likely that the strategy adopted for PREDICT - if it produced suitable input for the following steps - has been overfitted to work for rhodopsin. We strongly recommend to use another method or to replace the brute force algorithm applied afterwards by a more sophisticated one.

4.3 SCORING FUNCTION AND OPTIMIZATION PROCEDURE IN 3D

As mentioned above, the optimization algorithm of PREDICT is based on a simple brute force method. The underlying scoring function is adapted for the reduced protein representation developed by Herzyk and Hubbard[36] and is composed of two terms, membrane interactions as well as inter-helical interactions as stated in Eq. 5.

$$E = \sum_i E_{membrane}(Res_i) + \sum_{i,j} E_{int}(Res_i, Res_j) \tag{5}$$

In the following we discuss all formulas briefly. If given, we present the values for the parameters involved.

4.3.1 *Membrane interaction*

The first term of the PREDICT scoring function considers interactions of charged residues with the membrane. Let the membrane have a thickness of 30Å in z-direction, then one can define a favorable and unfavorable position for charged residues due to the bilayer property of a membrane:

$$E_{membrane}(Res_i) = \begin{cases} a < -1, & Res_i = \text{Arg, Lys, Asp, Glu and } 6\text{Å} < Z_i < 24\text{Å} \\ b > 1, & Res_i = \text{Arg, Lys and } (Z_i < 6\text{Å} \text{ or } Z_i > 24\text{Å}) \\ 1, & \text{otherwise} \end{cases} \quad (6)$$

Although Shacham et al. explain the importance of the location and orientation of charged residues in the membrane proteins, their membrane interaction term does not include this information. Instead, it uses only the z-coordinate of residues (most probably the C_α-atom) to assess their interactions with the membrane and does not depend on the exact side chain position, e. g., whether it points to the interior of the protein or to the membrane. Obviously, this term is not sufficient to cope with the different kinds of environments in a membrane bilayer.

4.3.2 *Inter-helical interactions*

The second term of the PREDICT energy function (see Eq. 5) evaluates the inter-helical interactions. The function is adapted for the reduced representation and cannot consider all details of a residue. For example, the ring in Tyrosin is represented by only one virtual atom, and hence inferring its exact orientation is impossible. The core part of this function is a distance-dependent function multiplied by several factors related to the properties of the underlying residue:

$$E_{int}(Res_i, Res_j) = \epsilon_{ij} \cdot \lambda_{arom} \cdot \lambda_{cat} \cdot \lambda_{polar} \cdot f_{ij} \quad (7)$$

The 4 factors are the Miyazawa and Jernigan contact energies (ϵ_{ij})[123] as well as specific energy contributions for aromatic, cation-π and polar interactions:

$$\lambda_{arom} = \begin{cases} \alpha > 1, & \text{i and j = Phe, Tyr, Trp, His} \\ 1, & \text{otherwise} \end{cases} \quad (8)$$

$$\lambda_{cat} = \begin{cases} \beta > 1, & \text{i = Arg, Lys and j = Phe, Tyr, Trp, His} \\ 1, & \text{otherwise} \end{cases} \quad (9)$$

$$\lambda_{polar} = \begin{cases} \gamma > 1, & \text{i = Arg, Lys, Asp, Asn, Gly, Gln} \\ 1, & \text{otherwise} \end{cases} \quad (10)$$

As these parameters do not change during optimization we can compute them in a preprocessing step to adapt Eq. 7 to

$$E_{int}(Res_i, Res_j) = \epsilon'_{ij} \cdot f_{ij} \tag{11}$$

where

$$\epsilon'_{ij} = \epsilon_{ij} \cdot \lambda_{arom} \cdot \lambda_{cat} \cdot \lambda_{polar} \tag{12}$$

and

$$f_{ij} = \sum_{k=1}^{M_i} \sum_{l=1}^{M_j} \frac{1}{M_i \cdot M_j} \cdot \left[a \cdot \left(\frac{R(C_{ik}) + R(C_{jl})}{D(C_{ik}, C_{jl})} \right)^m - b \cdot \left(\frac{R(C_{ik}) + R(C_{jl})}{D(C_{ik}, C_{jl})} \right)^n \right] \tag{13}$$

In Eq. 13 M_i and M_j are the number of atoms in residue i and j, $R(C_{ik})$ and $R(C_{jl})$ the radii of the current atoms and $D(C_{ik}, C_{jl})$ the corresponding distance between these atoms. The best results have been achieved using $m = 6$, $n = 4$, $a = 5$ and $b = 6$. Using this formula the contacts between the atoms of each residue are maximized, i. e., a optimal distance is computed based on the chosen parameters.

The question arises how to choose the parameters for λ_{arom}, λ_{cat} and λ_{polar}. In his diploma thesis, Thies showed that decoys can already be separated from crystal structures best when setting all of these parameters to 1.0.[124] This is a surprising result when considering that Shacham and co-workers suggested that these values have to be larger than 1.0. As a compromise, we decided to set these values to 1.1 for both reasons to be larger than 1.0 and simultaneously not to give too much weight to these kinds of interactions.

4.3.3 *Optimization methods*

Based on the results from the previous analysis, we adapted the 3D optimization procedure of PREDICT in two ways. First, we used more sophisticated methods, a Simulated Annealing as well as a Gradient-based algorithm, and second, to cope with the small optimization ranges, we restrained all rotational and translational degrees of freedom. In this section, we present the main features of the optimization methods and our restraint function.

Simulated Annealing
Simulated Annealing (SA) is a Monte Carlo method for minimization of a function where the global minimum can be far away from the current state. To do so, it accepts with some probability intermediate states, which are worse than the current one. The decision function used in SA is called *Metropolis criterium* (Eq. 14) and depends on three values: the energy difference between the current and the new state (ΔE), a temperature-like parameter T and a uniformly distributed random

number R in the interval [0,1]. A upwards step in the minimization is accepted if the Metropolis criteria is fulfilled.

$$e^{-\frac{\Delta E}{T}} > R \tag{14}$$

Including the temperature in the probability function has the desired effect that the probability to accept a worse state decreases when the temperature is lowered, i. e., when reaching the end of the optimization process. Different schedules to rescale the temperature have been proposed in the literature, e. g., a linear ($T_{n+1} = T_n - \alpha$) or geometric ($T_{n+1} = T_n \cdot \alpha$) one. We used the strategy first suggested by Lundy,[125] where α is a small number, in our case 0.001:

$$T_{n+1} = \frac{T_n}{1.0 + \alpha T_n} \tag{15}$$

Figure 4.8 illustrates a few temperature schedules and explains both the decision for the schedule and the choice for the value of α. On the one hand, α of the ge-

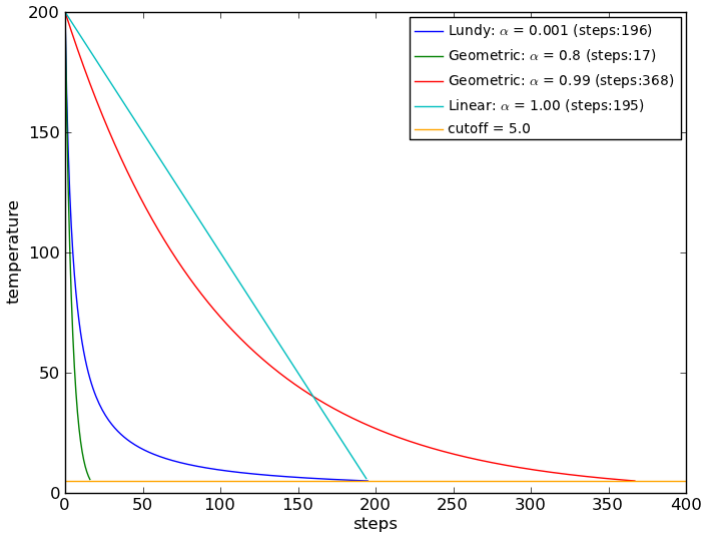


Figure 4.8: Exemplary selection of different temperature schedules. The choice of $\alpha = 0.001$ in the Lundy scheme can be inferred from the other schedules.

ometric schedule is chosen mostly between 0.8 and 0.99. Whereas the first value decreases the temperature very fast, the second one reduces the temperature only in small steps. On the other hand, setting $\alpha = 1.0$ in the linear schedule we need 196 steps when decreasing the temperature from $T_0 = 200.0$ to $T_{cutoff} = 5.0$. With our settings for the Lundy scheme, we combine the properties of all these schedules: a fast temperature decrease in the beginning, a slower one in the end of the optimization process and about the same number of steps compared to the linear temperature schedule. The slow decrease after a few steps helps also to avoid to stuck in one of many local minima.

Algorithm 1 Simulated Annealing

```
T ← T_0          //initialize temperature
S_c ← S_0        //initial state
S_b ← S_c        //save best state
E_c ← E(S_c)     //assess state
E_b ← E_c        //save best energy
while T > T_cutoff do
    t ← t_0      //initialize number of trials
    while t > 0 do
        S_n ← changeState(S_c)
        E_n ← E(S_n)
        if accept (E_n, E_c, T) then
            S_c ← S_n
            E_c ← E_n
            if E_c < E_b then
                S_b ← S_c
                E_b ← E_c
            end if
        end if
        t ← t − 1
    end while
    update(T)
end while
```

The general scheme (see Algorithm 1) of a SA has two loops. The outer loop decreases the temperature with regard to the predefined schedule until a cutoff is reached, while the inner loop specifies the number of trials that have to be performed at each temperature level. Because the helices are treated as rigid cylinders, we have six degrees of freedom, three for translation (x, y, z) and three for rotation (ψ, θ, ϕ), which are all updated in each single step. To change a state, a random number in a specified interval $[-\Delta_x,\Delta_x]$ is added to the current value of the corresponding degree of freedom. We set Δ_{trans} and Δ_{rot} to the very small values 10^{-2} and 8^{-5} (in radians), respectively. This was done to couple the step size to the current temperature by simple multiplication such that at the beginning (T_0 = 200.0), values up to 2Å for Δ_{trans} and 9.2° for Δ_{rot} can be chosen, while this range decreases to Δ_{trans} = 0.05Å and Δ_{rot} = 0.29° when reaching our temperature cutoff $T_{cutoff} = 5.0$.

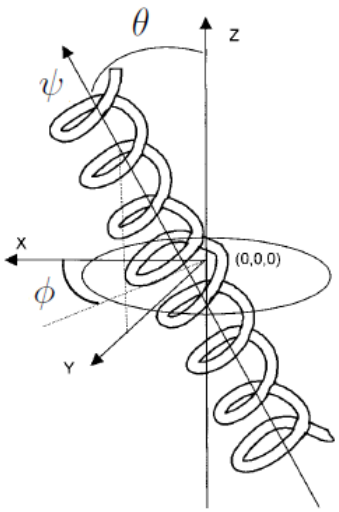


Figure 4.9: Rotation angles and axes.

Hence, a new state, e.g., for angle ϕ, is computed as follows, where the random number R is uniformly distributed in the interval [0,1]:

$$S_{n+1}(\phi) = S_n(\phi) + T_{n+1} \cdot (R \cdot 2.0 \cdot \Delta_{rot} - \Delta_{rot}) \tag{16}$$

Gradient-based optimization
Although we already decreased the step size during our simulated annealing procedure to be close to a local minimum, we applied afterwards a gradient-based optimization method to reach the closest optimum. There exist different local minimization techniques typically consisting of two steps: first, computing the descent direction and second, calculating the step width. We used the limited-memory BFGS (L-BFGS) algorithm, which is a variant of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method to update the Hessian matrix used in the optimization process.[126] This algorithm belongs to the class of quasi-Newton optimization methods where the Hessian (square matrix of second-order partial derivatives of a function) is only approximated and does not have to be computed directly. Here, we show only a general scheme (see Algorithm 2), and refer readers interested in local minimization techniques to the dissertation of Rurainski.[127]

Algorithm 2 BFGS method

$B_0 \leftarrow I$ //initialize Hessian with identity matrix
while $|\nabla f(x)| > \epsilon$ **do** //stop if norm falls below cutoff
  $P_k \leftarrow B_k^{-1} \nabla f(x_k)$ //compute direction
  Perform line search to obtain step size α_k
  $x_{k+1} \leftarrow x_k + \alpha_k P_k$ //change current state
  $s_k \leftarrow \alpha_k P_k$
  $y_k \leftarrow \nabla f(x_{k+1}) - f(x_k)$
  $B_{k+1} \leftarrow B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$ //approximate Hessian
end while

Side chain optimization
To rearrange the side chains in a proper way, Shacham and coworkers used a rotamer library. In our work, we used the backbone-independent rotamer library of Dunbrack and optimized the side chains in the SA procedure using a simple optimization method. For faster computation of new side chain positions we precomputed those in the preprocessing step and just selected one of those rotamers during optimization as follows: Before assessing a new state, we randomly rearranged several times up to three neighboring (in three-dimensions) side chains simultaneously and accepted the new positions only if the energy decreased. The neighborhood of a side chain is updated after each step that decreases the temperature.

Restraint function
While Shacham et al. optimized the decoys only in a small range, we used the following restraint function to penalize large movements away from the starting structure instead of a hard constraint:

$$C(x) = (1.0 + c_f \cdot |x - x_0|)^{2c_e} \tag{17}$$

Here, the exponent c_e and factor c_f are two parameters to adapt the strength of the restraint. We explain the choice of the values for each parameter directly when discussing our test cases in the next section.

4.3.4 *The 3D scoring function in practice*

In this section, we present the results of some simple test scenarios we used to analyze the applicability of our optimization procedure and, in particular, of the energy function of PREDICT. In each test case, we performed 50 times a simulated annealing run followed by our Gradient-based optimization procedure. We show the plots for 1U19 and 3ODU, while the others can be found in the Appendix.

Unrestrained crystal structure optimization
In our first test case, we used the crystal structures as starting conformation and optimized these without any restraint ($c_f = 0$ in Eq. 17). In this scenario, we can test if the energy function is consistent with the crystal structure or if structures with lower energies can be found far away. First, we checked to which extent the values

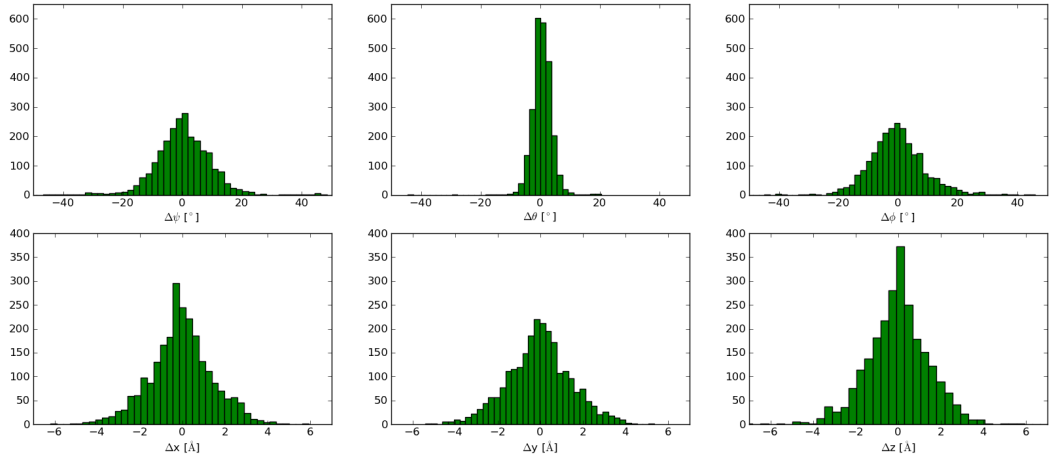


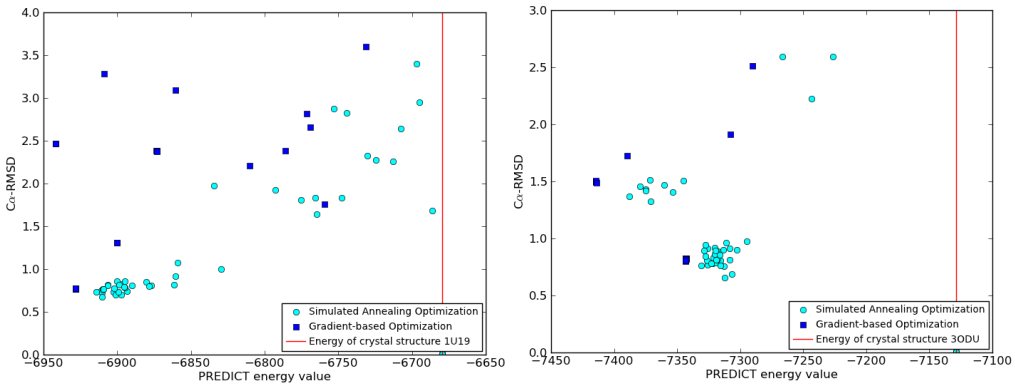
Figure 4.10: Histogram of change of degrees of freedom for X-ray optimization. The angles and translations are plotted for a better visualization in the interval [-50°,50°] and [-6Å,-6Å], respectively, although we obtained a few exceptions.

of the parameters changed during optimization. Therefore, we created a histogram (Figure 4.10) for each degree of freedom over all structures (7 structures $\times$ 7 helices

$\times$ 50 runs = 2450 values). We can see that angle ψ and ϕ changed significantly stronger than θ, but almost all of these values are in a proper range (<20°). Most of the helices moved by up to 2Å in different directions, but some of them even more than twice as far. This seems to be a large structural change, but one has to consider that a translation of all helices in the same direction would not change the structure at all. Hence, we need a further measure to compare two structures. A measure that is often used for this purpose is the root mean squared deviation (RMSD), where X and Y are the coordinates of two structures and n the corresponding number of atoms to be compared:

$$\mathrm{RMSD}(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|X_i - Y_i\|^2} \tag{18}$$

Since in the current state of the algorithm we are only interested in the backbone conformation, we used the C_α-RMSD where only the positions of the C_α-atoms of each amino acid are compared.



(a) 1U19 **(b)** 3ODU

Figure 4.11: 50 unrestrained simulated annealing runs each followed by gradient-based optimization applied to the crystal structure conformation.

Figure 4.11 illustrates the correlation between the PREDICT energy function and the C_α-RMSD of our optimized models and the corresponding crystal structures (start conformation) of 1U19 and 3ODU. For 1U19, we found one large cluster, with a C_α-RMSD of about 0.8Å, which seems to be a quite good result. But there are also three structures with an even lower energy (-6940) and a three times larger C_α-RMSD (2.48Å). Hence, if we do not use any restraint or - as PREDICT does - optimize not only in a small range, the PREDICT energy function would produce structures that are far away from a reasonable conformation. However, in those structures, the ϕ angle of H4 and H6 changed by about 130° and 85°, respectively. If we assume that the starting conformation is already close to the crystal structure, these models should not be generated due to the narrow range in which the decoys are optimized in the original PREDICT algorithm.

By and large, we obtained similar results for 3ODU. There are two clusters, one with a C_α-RMSD of 0.8Å and one with a C_α-RMSD of 1.5Å, while the latter has a lower energy. Here, there is only one degree of freedom that exceeds the allowed range, namely angle ϕ of H5, which changed by about 22°.
Summarizing these observations, we conclude that the PREDICT energy function seems to be too weak to discriminate between appropriate conformations and decoys. We found structures with high RMSDs, although we had a perfect starting conformation, including even the kinks in the helices. The search space can be restrained, but this implies that we need to be very close to the crystal structure with our start conformation, although we only use canonical helices. However, the start conformation prediction methods suggested by Shacham and coworkers are not sufficient, as we have demonstrated in Section 4.2.

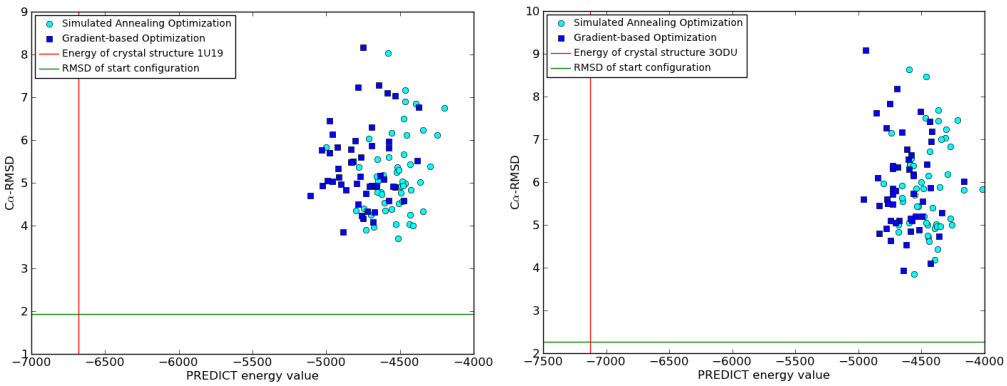
Unrestrained decoy optimization
In this test case, we replaced the helices in the crystal structures by their canonical variant to get the best possible start conformation according to the PREDICT approach. Table 4.5 shows the C_α-RMSD values between these models and the corresponding crystal structure. These values are quite good if we keep in mind that most of the helical structures are distorted.

Table 4.5: C_α-RMSD of the best possible decoy conformation.

	1U19	2RH1	2VT4	3EML	3ODU	3PBL	3RZE
C_α-RMSD	1.93	1.86	1.89	2.27	2.26	1.83	1.89

The C_α-RMSD between the crystal structures and a structure consisting of canonical helices is computed. Only the start conformation of 3EML and 3ODU have C_α-RMSD values above 2.0.

Again, we optimized these decoys using both methods (with $c_f = 0$ in Eq. 17) and plotted the C_α-RMSD in relation to the PREDICT energy function as shown in Figure 4.12.
Obviously, the models changed significantly stronger during the optimization process such that these results were useless for further modeling steps. For 1U19 and 3ODU the model with the lowest energy has a C_α-RMSD of 4.7Å and 5.6Å, respectively. To answer the question why the performance is so bad in this test scenario, we need to consider the main part of the PREDICT energy function given in Eq. 11. This formula assesses all kinds of interactions in a structure as good interactions, meaning that it also tries to maximize the number of contacts. In a GPCR, however, H3 is tilted in such a way that it opens a pocket for ligand binding and, therefore, induces a region without any inter-helical contact. This fact is not included in the simple scoring function. Hence, the tilt angle of H3 is decreased and all other helices are arranged around H3 to maximize the number of contacts. Here, it is even more obvious that we need to restrain the starting conformation of the models. Another observation that substantiates our suggestion is that we do not obtain any cluster due to important residue contacts. The PREDICT energy function is too weak and, in particular, the included membrane term is meaningless, as it takes only the z-



(a) 1U19 **(b)** 3ODU

Figure 4.12: 50 unrestrained simulated annealing runs, each followed by gradient-based optimization applied on the canonical variant of the crystal structure.

coordinate of the atoms into account. Hence, it does not matter how single helices are oriented. Moreover, we used canonical helices and these bad results is most probable also an effect of missing kinks.

Restrained decoy optimization

For the sake of completeness, we optimized the canonical variant of the crystal structures including our restraint function (see Eq. 17), where we set $c_e = 2$ and $c_f = 2$ for all translational degrees of freedom. The values for the rotational degrees of freedom were set to $c_e = 2$ and $c_f = 10$, respectively. Figure 4.13 illustrates that

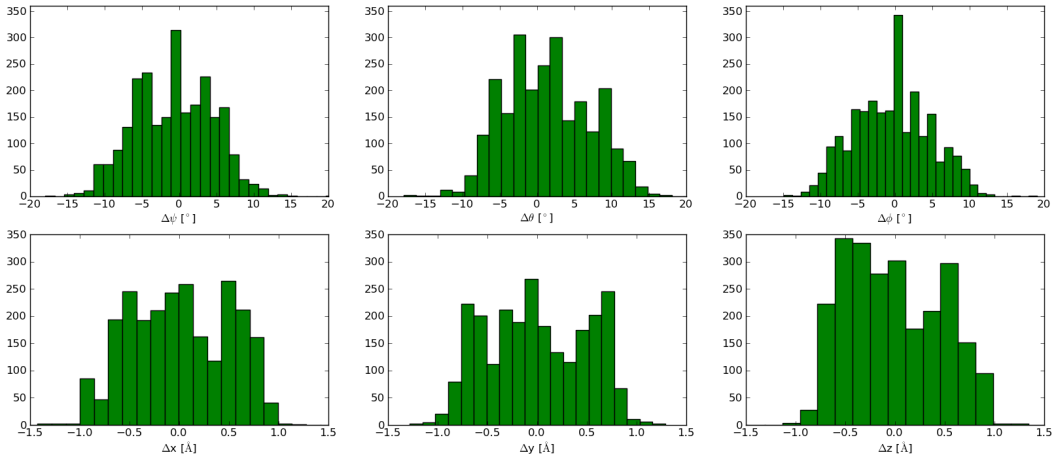
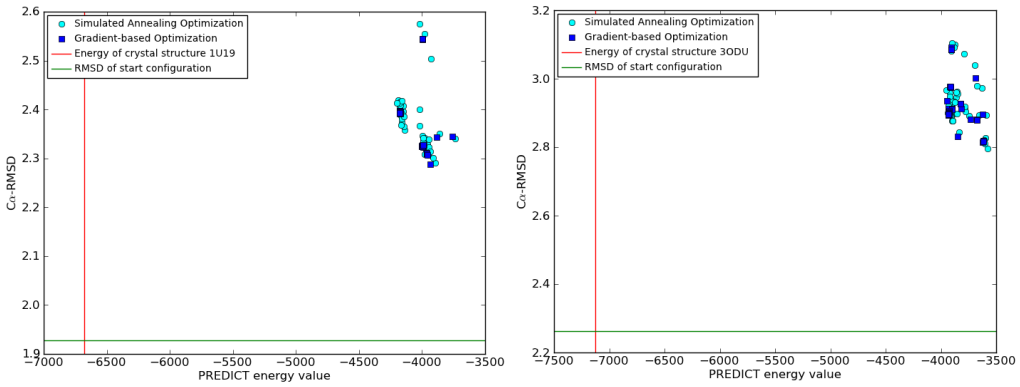


Figure 4.13: Histogram of change of degrees of freedom for canonical helix optimization. All deviations for the angles and translations are smaller than 20° and 1.5Å, respectively.

our restraint allows the rotational degrees of freedom to assume values that are

farther than 15° away from the start values only in few cases. This fits quite well to the brute force optimization procedure in a range of ±15° of PREDICT. Also, for the translational degrees of freedom we achieve proper values. Almost all of them are in the interval [-1.0Å, 1.0Å] with only a few exceptions up to [-1.5Å, 1.5Å]. Using our restraint, the C_α-RMSD increases only slightly by about 0.5Å (see Figure 4.14). But the question remains whether this is sufficient, when we consider that we started already from the best possible conformation. If we had some prediction errors in the 2D optimization procedure - and this is not unlikely -, we would achieve much worse results here. Hence, one cannot expect to get suitable models that can be used for further modeling steps. In the PREDICT procedure, the obtained models are filtered by experimental data and only the five best fitting models are taken to be converted to a full atomistic model and to introduce kinks using molecular dynamics simulations. In this last step, Shacham and coworkers simulated the whole structure for 280ps.



(a) 1U19 **(b)** 3ODU

Figure 4.14: 50 restrained simulated annealing runs each followed by gradient-based optimization applied on the canonical variant of the crystal structure.

4.3.5 *Conclusion*

The results of the last sections demonstrate that the PREDICT approach for ab initio GPCR modeling has many weaknesses. First, the TM prediction methods are not able to predict the helices in their full extent such that some binding pocket residues are missing. Second, the scoring function in 2D is mainly based on the hydrophobicity of amino acids. Here, none of the hydrophobicity scales guarantees that the helices are oriented in a sufficient way for further optimization steps. Third, the very simple 3D energy function maximizes the number of contacts but does not focus on a proper arrangement of the helices. This fact also explains why we were not able to find adequate values for the parameters λ_{arom}, λ_{cat} and λ_{polar}.

In our opinion, it is not promising to arrange canonical helices, regardless of the starting conformation. At this point, we stopped reimplementing the PREDICT

approach to check their general applicability on automated GPCR modeling. Although one step of the PREDICT approach is missing, the introduction of kinks, the results so far can never lead to an appropriate conformation. When arranging the helices in the reduced representation using the PREDICT energy function, they should already be kinked or we should at least allow kinks in this procedure. Hence, we adapted the algorithm of Shacham et al. by using MD simulations for individual helices. A short methodological background, the difficulties of simulating single helices, and the results are discussed in the following section.

4.4 ASSIGNING KINKS IN HELICES

The last and also an important step in the PREDICT procedure is the introduction of kinks in helices. So far, the helices are optimized in canonical form, but helices are often distorted. To cope with this challenge, Shacham et al. used the technique of Molecular Dynamics (MD) simulation, which is very useful to analyze the movement of molecules.[128] To this end, Shacham and co-workers added explicit water to the binding pocket and simulated the structure 280ps. This allowed helices to form kinks without changing the fold of the whole model significantly. However, the models obtained so far by the PREDICT algorithm feature, even in the best case, too high C_α-RMSD values to use only a short simulation in the final step.
Due to these results, we did not use MD simulations according to the PREDICT procedure. We applied this technique on individual helices instead, to achieve helical structures that are more similar to those in the crystal structures than their canonical variant and to arrange these helices afterwards.

4.4.1 *MD Simulation Setup*

Unfortunately, helical structures are not very stable if simulating them in pure water.[129] To be able to simulate helices for a time span longer than 280ps without a strong dissociation and to simulate the interaction with the hydrophobic part of the membrane, we also added trifluoroethanol (TFE) in addition to the water-molecules. In a former study, van Buuren et al. showed a stabilizing effect of TFE on helical structures (see Figure 4.15).[129] According to this study, the helices were solvated in a 30%(v/v) water/trifluoroethanol-mixture.

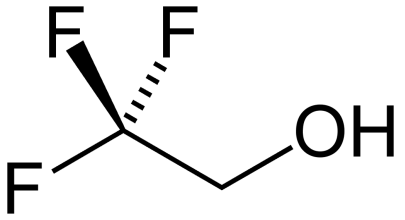


Figure 4.15: 2,2,2-Trifluoroethanol

For our simulation, we used the software package Gromacs, which is designed to simulate Newton's equations for millions of particles, in particular, for biochemical molecules.[130] In addition, it provides tools to analyze the trajectory of the system. Next, we give all parameters needed to reproduce our results.
The helical peptide was created using the class *PeptideBuilder* in BALL. Since the helical ends are charged, we added ACE/NME-caps to these ends applying the

PeptideCapProcessor for neutralization. The first step in the preprocessing was to add a cubic (*editconf -d 1.1*) box to the peptide, and then to convert the pdb to gmx format (*pdb2gmx*), where we set the Amber99SB forcefield (*-ff amber99sb*) and the SPC water model (*-water SPC*). Applying *genbox*, we solvated the helix with our water/TFE-mixture (*-cs tfewater.pdb*). Finally, the system was neutralized by replacing water molecules with the adequate number of counter ions (*genion -nn/-cp*). For the minimization and equilibration step, we used the *genrestr* script to restrain the backbone of the peptide.

Minimization

Table 4.6: Parameters for minimization

name	value	description
integrator	steep	Steepest descent algorithm for energy minimization
nsteps	1000	Number of steps
coulombtype	PME	Fast Particle-Mesh Ewald electrostatics

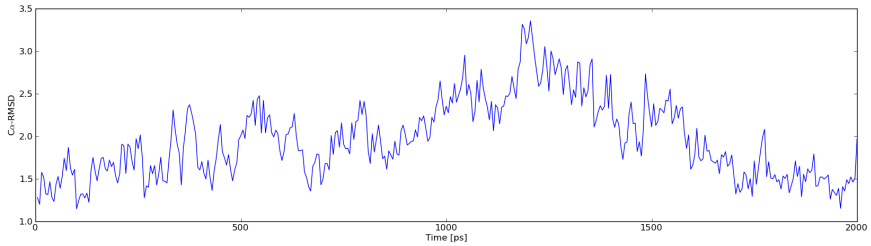
The backbone-restraint peptide was minimized by a steepest descent algorithm for 1000 steps. The coulomb type was set to PME.

Equilibration / Simulation

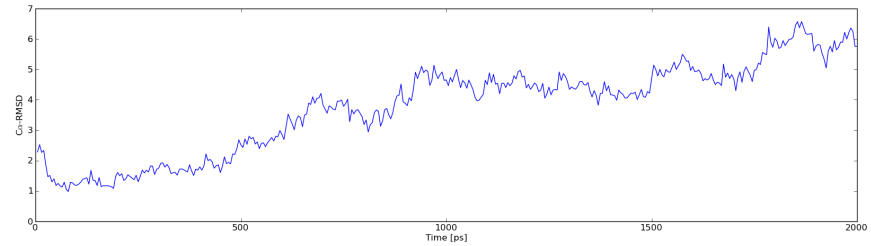
Table 4.7: Parameters for equilibration and simulation

name	value	description
integrator	md	Leap-Frog algorithm
pbc	xyz	Periodic boundary condition in all directions
coulombtype	PME	Fast Particle-Mesh Ewald electrostatics
rcoulomb	1.5	Distance for the Coulomb cut-off
vdw-type	switch	Use switching for VdW
vdw-switch	1.2	Where to start switching the LJ potential
rvdw	1.4	Distance for the LJ cut-off
tcoupl	v-rescale	Temperature coupling using velocity rescaling
tau_t	0.1	Time constant for coupling
ref_t	310	Reference temperature for coupling
pcoupl	parrinello-rahman	Extended-ensemble pressure coupling
pcoupletype	isotropic	Isotropic pressure coupling
tau_p	2.0	Time constant for coupling
ref_p	1.0	Reference pressure for coupling

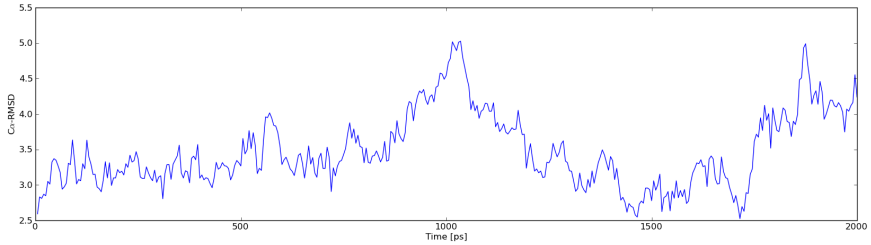
The backbone-restraint equilibration was performed for 5ns to give the TFE enough time to arrange around the peptide. Afterwards we did a 2ns simulation without any restraints. The parameters specified in this table were used for both.



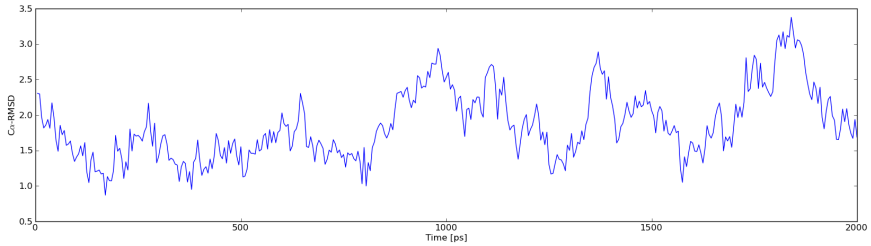
(a) H1 of 1U19



(b) H6 of 1U19



(c) H4 of 3ODU



(d) H6 of 3ODU

Figure 4.16: Trajectories of 4 peptides in canonical form. Only H6 of 1U19 dissociates, while the others are more stable for a longer time.

4.4.2 *Results and Discussion*

After extracting the helices from the system, we focused on their stability. Figure 4.16 illustrates the C_α-RMSD values between the structures in the trajectory and the one in the crystal structure sorted by time. We see that H1 of 1U19 as well as H4 and H6 of 3ODU are very stable during the whole simulation. This demonstrates the effectivity of TFE regarding the stability of helices. Unfortunately, there are some extreme exceptions like H6 of 1U19. From the very beginning, this helix started to dissociate such that finally a nonrealistic model was obtained. However, in the first 500ps, the structure has low C_α-RMSD values between 1.0Å-1.7Å, which is quite good compared to its canonical variant.

To check to which extent the helices from the MD simulation can improve our GPCR models compared to canonical ones, we computed the C_α-RMSD when mapping the best model from the MD onto the crystal structure (see Table 4.8). Only for H4 of 3ODU we were not able to produce an adequate structure, while in other cases, for instance, H6 of both structures, the models obtained are twice as good as the canonical variant. When mapping the helices to those of the corresponding crystal structure, the TM region can be modeled in the best case with a 40% lower RMSD value.

Table 4.8: C_α-RMSD of the best helix produced by MD

	H1	H2	H3	H4	H5	H6	H7	TM region
1U19 - Canonical	1.15	2.26	1.11	1.18	2.24	2.29	2.30	1.93
1U19 - Simulated	1.15	1.72	1.07	0.56	1.69	0.98	1.43	1.28
3ODU - Canonical	1.92	1.88	1.35	2.66	2.60	2.39	1.92	2.26
3ODU - Simulated	1.55	0.89	0.79	2.53	1.81	0.87	1.21	1.49

All of the best single helices obtained from the MD are better than their corresponding canonical variants. Only H4 of 3ODU has a C_α-RMSD higher than 2.0Å. The RMSD of the TM region is reduced by up to 40%.

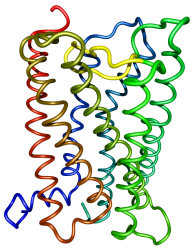
In this case, the main challenge is to extract the proper snapshots from the trajectory. In his recent Bachelor's thesis Lund showed that his methods are able to filter all meaningless snapshots easily.[131] However, although the best structure belongs in almost all cases to the top 5 ranked structures, it is only in a few cases best ranked. Moreover, the best ranked structure is sometimes worse than the canonical variant. But even if we are able to extract the best helical structure from the trajectory, it can only be used as a starting point in the modeling procedure regarding its kink. The reason might be that in our MD simulations inter-helical interactions that play a decisive role in folding processes are neglected.
Since we are interested in fully automated modeling, we did not go further in this direction, i. e., we did not analyze the impact of the best obtained helices from MD simulation with regard to the quality of our models as we were not able to extract the suitable snapshots automatically.

4.5 CONCLUSION

To conclude the study of reimplementing the GPCR modeling approach of Shacham et al., we have to emphasize several points. First, we demonstrated that the generation of suitable initial models completely failed. Neither for the new available GPCRs nor for bovine rhodopsin we were able to reproduce the results obtained by the authors. The methods they suggested, e. g., for prediction of helical regions or the initial orientation of those, are very error-prone and could only be manually improved in few cases. Nevertheless, we focus on automated GPCR modeling and hence do not want to rely on manually refinements.
Second, we clearly showed that the energy functions works more or less well when applying the optimization procedure on native structures. However, Shacham et al. used canonical helices and here, we obtained a dramatically decrease in the quality of the models. Thus, the energy function might be useful for small refinements but it is for sure not sufficient to optimize decoys at all.
Third, we tried to improve the PREDICT approach by using already kinked helices for generating suitable initial models. Therefore, we explored if molecular dynamics simulation can provide us more information about distortions in helices, i. e., where and how strong the helices are kinked. With our setup we were able to get in almost all cases improved helical structures during the simulation, however, to extract the proper snapshots could not be done in an automated fashion.

The main challenge when modeling the backbone of a GPCR is modeling the details, in particular, the kinks in the helices in a proper way. The idea we had is to treat helices as a combination of several cylinder fragments instead of a single rigid body. Thus, the helices can form a different shape during the optimization process. In this way, kinks are modeled with respect to their tertiary interactions. The only ingredient we need for this process to work is the position of the hinges. In a recent project, we predicted kinks from the sequence using string kernels for support vector machines. This study is described in the next chapter, while in Chapter 6 we examined how the information gained can help us to model GPCRs.

5 *Kink Prediction*



The prediction of structural elements based on protein sequences is a major task in bioinformatics. Consequently, many algorithms dealing with secondary structure prediction have been developed, starting with very simple methods in the 1970s to more complex ones in the 1990s. Some of the previously reported procedures rely on templates,[132,133] applying machine learning techniques,[134–136] or combining both.[137] For a more detailed review, the interested reader is referred to Pirovano et al.[138]

However, it becomes increasingly apparent that distortions of perfect geometries in secondary structure elements are very important to create structural diversity from simple building blocks, e. g., in helix bundle membrane proteins.[139] The distortions can be divided into different types, e. g., wide turns or kinks in helices. Especially the latter one, which changes the helical axis noticeably and rather abruptly, yields a significant change of the structure. But even though the knowledge about kinks is crucial for successful modeling of new structures, the number of available algorithms for computational kink prediction is relatively small.

In 2003, a promising sequence pattern descriptors approach was developed by Rigoutsos et al.[140] Based on motifs extracted from 17 proteins, the authors created a search engine to discriminate not only between ideal helices and distortions but also between different distortion types from perfect α-helicity. However, the low prediction accuracy for new sequences not included in their data set (e. g., new GPCRs) as well as their very low false positive rate (0.03%) for nonmembrane spanning helical structures and nonhelical region indicates an overfitting of their descriptors. An alternative approach, due to Yohannan et al., is based on the so-called evolutionary hypothesis for kink generation.[139] Information derived from homologous protein sequences can be used for kink prediction. In their study, the authors focused on kink patterns in different G-protein coupled receptor (GPCR) classes and also on eight unrelated membrane structures. For these proteins, they predicted 36 of 39 proline and 14 of 17 nonproline kinks correctly without any false positives. While the former approaches work on the sequence level and are relatively fast, Hall et al. used a complex molecular dynamics simulation setup to reproduce kinks in 405 helices of which 44% have been kinked.[141] Approximately 79% of the 62 proline-induced kinks were predicted correctly with a very high specificity. Interestingly, the prediction accuracy of this structural approach decreased to 58% and 18% for vestigial proline and nonproline induced kinks, respectively. In 2010, Langelaan and co-workers provided a large data set of 842 TM helices. These helices have been automatically annotated using the MCHELAN algorithm, which found a kink in 64% of all cases. Thereafter, they applied support vector machines to predict kinks in a range of ± 4 residues.[142] The approach achieved a maximal accuracy

of 74% when predicting based on the presence of proline. However, the F-scores of the prediction results were never above 0.6. This relatively weak score might indicate one of three things (or a combination thereof): either the sequence is only one (small) factor in producing kinks, the data set used was too error prone due to automated annotation, or the applied kernel functions are insufficient for predicting kinks.
Quite recently, Bowie et al. published an approach called TMKink.[143] They used a neural network with 5 hidden nodes and achieved the best performance for their data set (323 kinked and 567 nonkinked helices) for a window size of 9, resulting in a sensitivity and specificity of 0.7 and 0.89, respectively.

The approach put forth in the present work has been published in the Journal of Chemical Information and Modeling in 2011.[144] It has been developed independently from the work of Langelaan et al. but is similar in spirit. We also started to collect a large data set, but instead of relying purely on automated kink annotation, we created a manually curated data set. We examined the sequential neighborhood of the annotated kinked residues and computed the solvent accessible surface area (SASA) per residue for kinked and nonkinked helices to quantify the influence of neighboring amino acids and helices.
In addition, to accentuate the need of our manually annotated data set, we used three alternative state-of-the-art methods for automatic kink annotation from the three-dimensional structure and compared their respective qualities. All four data sets were used to train string kernel-based support vector machines for predicting kinks from the protein sequence alone. Furthermore, we compared our performance to TMKink to show the significant better performance of string kernels for support vector machines. In summary, our main goal in this work is to create a highly accurate data set for kink prediction, compare its quality to data sets automatically derived from the three-dimensional structure, and apply a statistical learning method to predict kinks from the primary sequence. This will yield insights into the state-of-the-art on structurally based kink detection, on the influence of annotation errors on statistical predictors, and on the influence of different sequence features on the likelihood of kink formation. In its current state, the method does not address the problem of determining the exact kink position in a distorted helix, but preliminary work in this direction will do.

5.1 MATERIALS AND METHODS

5.1.1 *Data Set Generation*

To create a reliable data set, we used the database MPtopo[3] to obtain all currently (January 2011) available α-helical transmembrane (TM) proteins, extracted their helices, and removed the ambiguous ones (pairwise sequence identity >95%) using the PISCES algorithm.[145] For a higher accuracy, the extraction of a single helix was manually curated by visual inspection using BALLView.[87] To select only those helices whose largest part is inside the membrane, we applied the online tool TMDet.[146] The final data set contains 132 proteins including 1014 helices (see Ap-

pendix). The largest helix has a length of 43 and the smallest of 12 residues, indicating at least 3 turns.

5.1.2 *Kink Definition*

Given this data set, the next task was to classify these helices as kinked or nonkinked. Previous work has relied on automated kink detection from three-dimensional structures.[142] From our own experiences, however, we expect the automated techniques to fail in some instances, which is also reflected in the fact that there exists no consistent definition of a kink. Moreover, in a canonical helix, the backbone torsion angles ϕ and ψ are fixed at -57° and -47°, but according to the literature, real-world helices are commonly slightly curved toward the solvent, yielding torsions of about -62° and -41° with a higher variance, respectively.[147] Therefore, many helices are hard to classify using simple methods based on large local deviations of torsion angles; in our brief analysis, deviations for kinked, nonkinked, and curved helices look sometimes very similar. In addition, a residue causing merely a wide turn can lead to the same magnitude in angle deviation as one inducing a kink.
To quantify the quality of automated structural kink detection and to understand the influence of annotation errors on statistical learning schemes, we decided to build a hand-curated data set first. To this end, we used visual inspection using BALLView to determine whether a given helix contains a kink. Exact criteria for the manual annotation of a kink are very hard to define. Each helix has to be examined from several perspectives to identify the residue producing a kink. Indeed, there are some arguable cases where different viewers will report different answers, but this problem is similar to the one of finding an appropriate cutoff in automated methods. Another advantage of manually annotated kinks is the possibility to have a look at local changes and their global effects at the same time. To reduce the bias of only one viewer, the helices were checked by two people. As a rule of thumb, a kink can be an abrupt change of the helical axis or a twisted residue causing for example a shifted axis. Small increases of the helix diameter for one turn, known as wide turns, were not annotated as kinks. While this method is also not entirely free of errors, it is more reliable than any automated criterion we tested. Our annotation leads to manually annotated data set (MDS) with 367 kinked, 461 nonkinked, and 196 curved helices (see Figure 5.2).

We then compared the data set to two automatically generated ones. First, we adapted principal component analysis (PCA) to compute the helical axis from the backbone atoms.[148] To this end, we split a helix into two segments, where the smallest one consisted of at least five consecutive residues, and computed the axis, i.e., the first principal component, for both parts. The axes had to be slightly adjusted - one end was set to the shortest distance point of the two vectors - to obtain a continuous axis (Figure 5.1). Afterward, we computed the minimal distance m over all backbone atoms to the computed axis. This procedure was done for all possible pairs of two segments, and we chose the pair yielding the best fit to the original helix, which is the one with the maximal value m. Finally, we classified a helix as kinked if the angle was larger than a predefined cutoff and defined the residue closest in structure as the kink position. Supposing that our manually curated data set is indeed the most reliable one, we chose the cutoff such that the proportion of kinked and nonkinked helices was similar to the one of MDS. Therefore, an angle value of at least 11° was used, yielding to 364 kinked helices in the data set called PCA data set (PDS). Note that this method can only find global effects of disrupted helices and, therefore, works for at most one kink per helix.

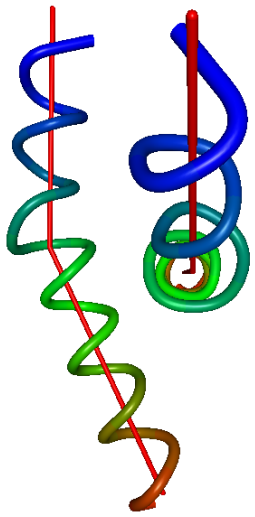


Figure 5.1: Side and top views of helix 1 of bovine rhodopsin (PDB ID: 1U19) with its computed helical axes using PCA (red lines).

In a second annotation approach, we applied the method HELANAL[149] using the python toolkit MDAnalysis.[150] In this method a helix was defined as kinked, if at least one local bending angle is larger than 20°. Although a few residues (mostly in a row) might fulfill this property, we set the kink to the position of the largest bending angle. In contrast to our PCA method analyzing global effects, this algorithm finds only local ones because just seven residues are used for the computation of an axis. Applying this method we obtained the data set HELANAL with 303 kinked helices.

Based on the choice of annotation algorithms for our manually extracted helices from α-helical membrane proteins, we can compare the performance of our string kernels on a data set created on both manual and automated methods, relying either on global or local effects.

Last but not least, we converted the data set available on the MC-HELAN Web site to our format. We suppose that the MC-HELAN method is currently the best available automated kink annotation method and, hence, an upper bound for our comparison between manually and automatically annotated kink data sets. Note, in contrast to our manually created data set (or to others published so far), the one created by Langelaan and co-workers has an inversed number of kinked and nonkinked helices.

When applying the TMKink web server on all four data sets, we defined a helix as kinked if at least one kink was predicted, since we focus only on the classification of kinked and canonical helices. The cutoff was adjusted to $t > 0.7$ to obtain a higher balanced accuracy and, hence, to be able to compare our performance with the best possible results we can achieve using TMKink.

5.1.3 *Statistical Methods*

Support vector machines (SVMs) using string kernels have been applied to all data sets using a nested five-fold cross validation setup.[151] The main idea of string kernels is to compare strings by means of the substrings they contain, while these substrings are not assumed to be contiguous. Whereas SVMs have previously been used for kink detection,[142] the use of string kernels as a measure of similarity is novel to this field. In addition, string kernel SVMs have been shown to produce clearly superior results to classical SVMs in fields, such as protein classification,[152–154] prediction of t-cell epitopes,[155] and other structural biology problems. Whereas many different string kernels have been proposed in the literature,[156] we concentrated our approach on the following:

Alignment Kernel
The alignment kernel is strongly motivated by the Needleman-Wunsch alignment score. We used the simple edit distance as scoring function to allow mismatches, insertions, and deletions of amino acids in the input sequences. Hence, the kernel value is the minimum number of operations to transform one sequence into the other.[157]

K-mer Kernel
In general, a k-mer kernel measures sequence similarity by shared occurrences of fixed-length patterns in the data, allowing for mutations between patterns (mismatch kernel) and a weighting of pattern frequencies (spectrum kernel). In this work, we also allowed a combination of both as well as a combined spectrum kernel of different k-mer sizes.

For the sake of convenience, we introduce the following nomenclature: K for the k-mer length, M for the number of allowed mismatches, and W1(0) to (not) weight the multiple occurrences of the k-mers. Thus, K123_M0_W1 takes all k-mers of size 1-3 where no mismatches are allowed and the occurrences are weighted. In the case of the alignment kernel, we tested different values for the parameter γ, e. g., Alignment_G0_01 represents the alignment kernel with a chosen γ of 0.01.
The SVM setup was realized using libsvm[158] with the precomputed kernel option and was integrated into the Biochemical Algorithms Library BALL.[159] The parameter C of the SVM was determined using five-fold nested cross-validation (from 10^{-6} to 10^{3}).[160]
At this point we want to stress that finding an optimal setup for training the SVM classifiers was out of scope of this work. Indeed, we assume that the prediction

accuracy can be further increased with a more sophisticated setup such that our results can be seen as a lower bound for this method. This will be a focus of future work.

Statistical Performance Measures
To obtain an impression of the classifier's performance, we determined the so-called confusion matrix relating true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).[161] Based on this matrix, several performance measures, such as specificity, sensitivity, and accuracy, can be easily calculated. Due to the imbalance in our data sets, we will specify the balanced accuracy and the F-score. The significance value for the solvent accessible surface area analysis was calculated applying the Welch two sample t-test using R.[89] In this case, the null hypothesis for the two-sample t-test was $\mu_1 = \mu_2$. Hence, the smaller the computed value, the more convincing the rejection of the null hypothesis.

5.2 RESULTS AND DISCUSSION

5.2.1 *Data Set Analysis*

As a first step of our analysis, we studied the manually annotated data set in detail. Figure 2 shows the length distribution of all helices: 461 helices were defined as nonkinked, with a maximal frequency at the length of 20 amino acids, equaling the number of residues needed to completely cross a typical cell membrane. In 357 cases, we found a residue obviously causing a change of the helical axis, thus introducing a kink. These helices are more uniformly distributed between 19 and 32 residues and, as expected, tend to be longer than canonical helices (the kink allows to fit a longer helix into the membrane). Altogether, there are 196 helices featuring a distortion that could not be exactly assigned to one residue due to a curved structure. We removed these helices in the following to reduce the noise such that finally 357 kinked and 461 nonkinked helices are included in our manually created data set (MDS).
To obtain information about the data set and the amino acids it contains, we calculated their percentage distribution in the whole protein (see Supporting Information) and the helical regions as well as the determined kink positions (see Table 5.1).
Two interesting groups of amino acids are immediately apparent: the first and most important one for our work contains glycine, serine, and proline (marked yellow). Their percentage occurrence value is smallest when considering the helical region. The special role of proline and glycine is well established from many other studies,[162,163] and serine is also known as a potential helix breaker.[141] Thus, statistical analysis of the amino acid distribution at kink positions in the data set confirms our manual annotation of the helices.
Another interestingly distributed group, marked in orange, contains both acidic amino acids (aspartic and glutamic acids) and two basic ones (histidine and arginine) as well as asparagine. The members of this group appear only very rarely at kink positions (in total 4.8%), even though their general occurrences in proteins

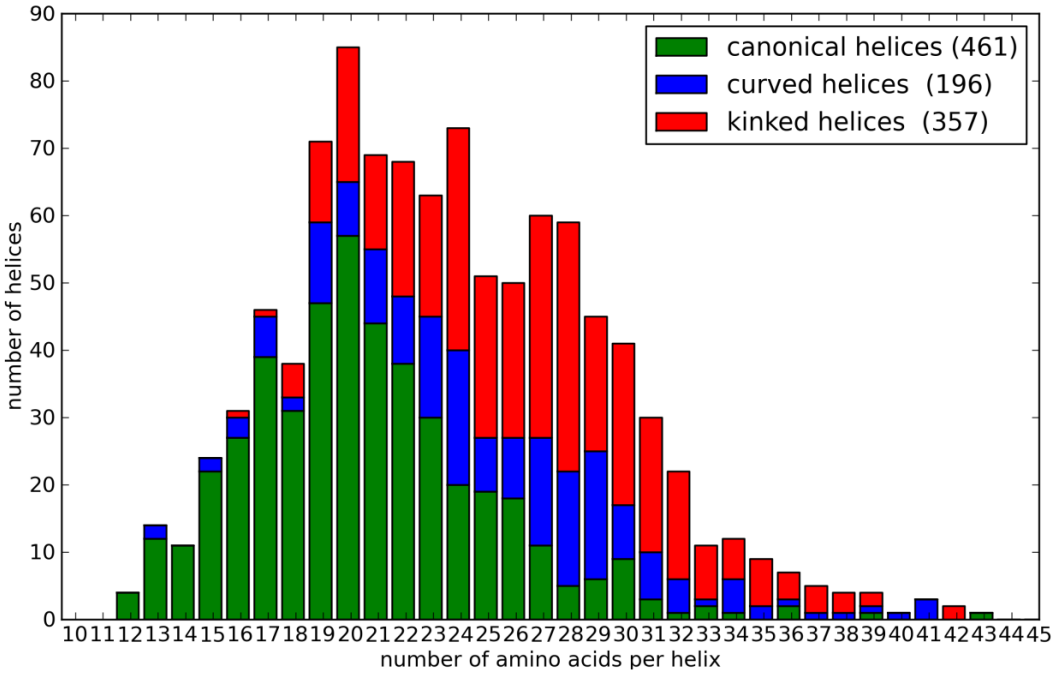


Figure 5.2: Length distribution of all helices in our data set. The ratio of kinked to canonical helices swaps if the length exceeds 24 amino acids.

Table 5.1: Amino acid distribution in proteins, helices and at kink positions

AA	Protein	Helix	Kink	AA	Protein	Helix	Kink
A	9.5	11.6	9.2	M	2.9	3.6	4.6
C	1.1	1.3	1.1	N	3.2	1.7	1.1
D	3.3	1.1	0.0	P	4.6	1.9	6.2
E	3.9	1.8	1.5	Q	2.7	1.8	2.4
F	6.4	8.1	8.1	R	3.9	2.4	1.1
G	8.5	8.2	10.5	S	5.7	4.8	6.5
H	2.0	1.6	1.1	T	5.5	5.5	4.3
I	7.6	10.3	7.8	V	8.3	10.4	11.6
K	3.4	1.8	1.9	W	2.4	2.8	2.4
L	11.6	15.8	15.1	Y	3.5	3.5	3.5

Percentage distribution of all amino acids in the whole protein, the extracted helices and at the manually annotated kink position of MDS. The two interesting groups are highlighted in orange (low occurrences at kink position) and yellow (high occurrences at kink position).

and helices are much higher (in total 16.3% and 8.6%, respectively). In addition, lysine, the third basic amino acid, also appears unfrequently at a kink position (1.9%). The rare occurrence of these amino acids at kink positions is due their sequence position in the helix. They appear mostly at the end of a helix (because of the membrane environment), where only a few kinks are determined (altogether 43 kinks in the first and last 30% of a helix). Hence, these findings do not allow to draw simple conclusions on the relevance of this group for kink formation. Besides the kink

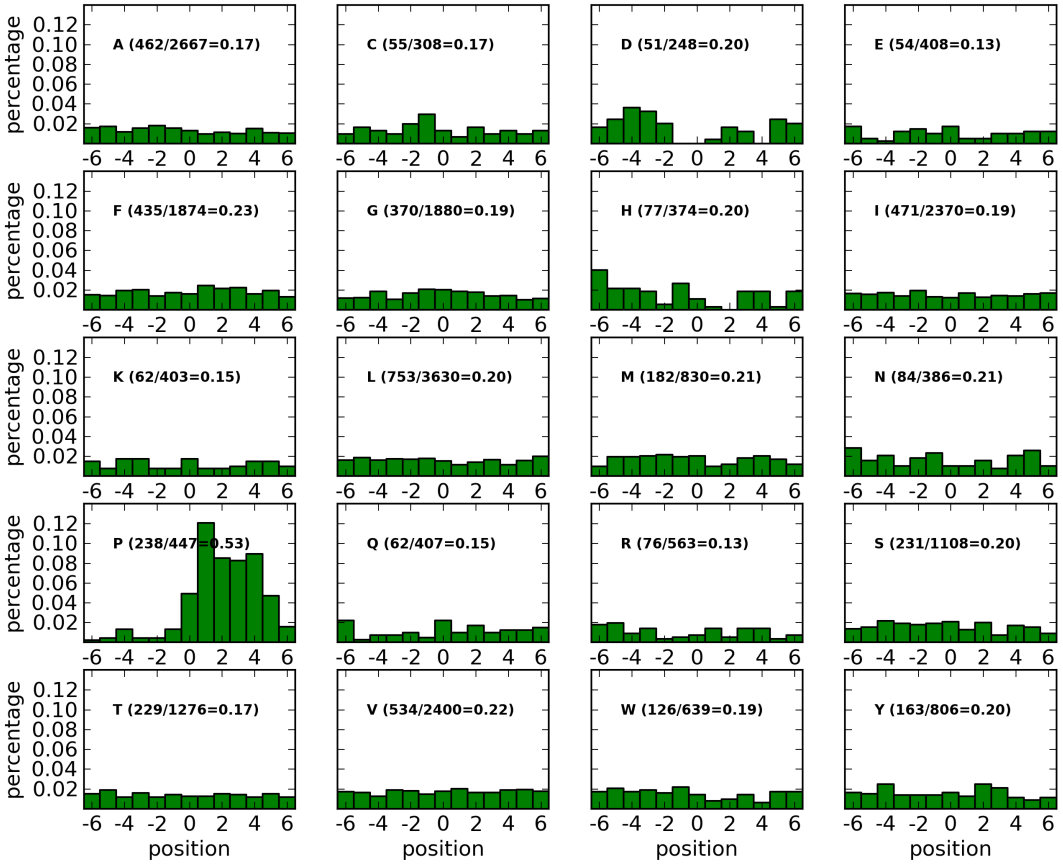


Figure 5.3: Neighborhood of kinks. For every amino acid type, we computed the probability to be found at a specific position ($\pm$ 6 residues) next to a kink (position 0). We divided the number of each amino acid found in this region by the number found in the complete helix. The expected value is 20.5%.

position itself, the kink environment may play an important role. Therefore, similar to Langelaan et al.,[142] we computed the occurrence probability for each amino acid around an identified kink (Figure 5.3). There are four amino acids with an under-representation of at least 5% in this region: arginine, glutamine, glutamic acid, and lysine, which supports the results drawn from Table 5.1. Over 50% of all prolines in all helical sequences occur in a range of $\pm$5 residues around a kink, in particular, between 0 and +5 amino acids next to a kink, which corroborates its great relevance. For this reason, we defined all kinks as proline induced if a proline occurs in this range. This is slightly different to the annotation results of the MC-

HELAN algorithm,[142] where especially positions 2 and 3 after an identified kink are over-represented by proline. Glycine and serine do not reveal such a distribution, because their ratio between helix and kink occurrences is much lower compared to proline. Some amino acids were not found at specific positions, e. g., neither was a histidine present one or two residues after a kink nor an aspartic acid found at residue position -1, 0, or 4. Another aspect is the small increase of aspartic acid one turn before a kink, which fits very well to the data of Langelaan.

Table 5.2: Amino acid composition in kinked and canonical helices

	-4	-3	-2	-1	0	+1	+2	+3	+4
A						0.44	0.49		
C						0.37		0.39	
D	11.9	10.6	6.64	0.26		0.44			0.33
E	0.14				3.09		0.44		
F									
G									
H		2.32			5.31	0.44	0.44	3.98	
I									
K	7.97	9.29		2.65	9.29			3.98	2.65
L									
M							0.44	2.46	
N	2.32			11.9					
P		0.29			23.9	71.7	49.1	23.2	25.2
Q	0.49				3.54				
R		3.54	2.65			9.29		2.12	
S								0.44	2.17
T									
V									
W			2.21	4.98					0.44
Y	2.21				2.27	2.21	5.31	2.25	

Ratio of frequencies in kinked and canonical helices is shown. Numbers are given if a residue is at least two-fold over-represented (>2.0) or under-represented (<0.5) in kinked helices compared to nonkinked helices. Position 0 refers to the annotated kink and the center of the helix in kinked and canonical helices, respectively.

In our last data set analysis, we compared the amino acid composition of kinked and canonical helices at specific positions. To this end, we superimposed the kink position of the kinked helices on the center of nonkinked helices, computed the occurrences of each amino acid in a range of ± 4 residues, and calculated the ratio of their frequencies. As illustrated in Figure 5.2, several amino acids are over- or under-represented at different positions. Here, we want to mention only three: the five times over-represented tryptophan (it has two rings), the several times over-represented arginine and lysine (long and charged side chain), and the charged aspartic acid, which is over-represented at the first three positions but under-

represented at positions four, six, and nine. To determine the exact kink position in later projects or with regard to a better understanding of kink formation, this analysis might be very helpful.
To explore the influence of neighboring helices, we calculated the SASA per residue for each helix (see Figure 5.4). We applied a two-tailed t-test to assess whether the SASA means are statistically different for kinked vs ideal helices, for kinked vs curved helices, and for ideal vs curved helices. For the first comparison (kinked against the nonkinked helices), we obtained a p-value of 0.0031, while the p-value for the second case (kinked against curved) was 0.0511. The first value in particular indicates a large difference in the environment of the compared helix types. The environment for nonkinked and curved helices, however, seems to be very similar (p-value: 0.8354). Due to the significantly smaller SASA mean value for kinked helices, we conclude that the resulting larger amount of potential tertiary interactions can help to enforce and stabilize these kinks.

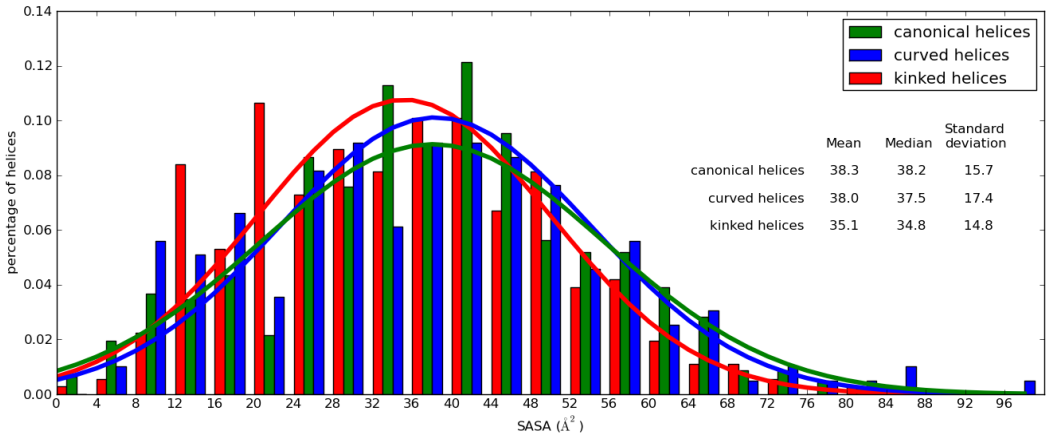


Figure 5.4: Histogram of the SASA for all identified helices. In addition, the corresponding probability distributions are shown.

5.2.2 *Evaluation of the Automated Detection Methods*

Figure 5.5 shows a Venn diagram for the kinked helices. Only 59% of all manually annotated kinked helices have been identified by both automated methods, but in total, only 29 kinks have been manually, but not automatically, detected, and 125 kinks were annotated automatically by either PCA or HELANAL, but only 4 of these were detected by both methods. As mentioned above, these methods focus on different aspects of structural changes in kinked helices and are insufficient to create a reliable data set on their own.

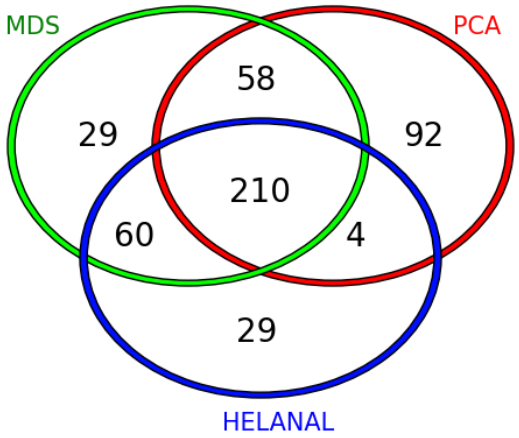


Figure 5.5: Venn diagram for kinked helices of our 3 data sets: 210 helices are defined as kinked by all the methods, and 29 have been identified by the two viewers as kinked but neither by PCA nor by the HELANAL method, while 125 sequences are identified by at least 1 automated method, they have only 4 of these in common.

Table 5.3 shows if the manually and automated methods defined a helix as kinked, these positions do not differ much. Over 90% are within 1 helical turn (4 residues). This means that automated methods work well, but our further results will demonstrate that the exact kink annotation is necessary for predicting kinked helices with a high accuracy.

Table 5.3: Comparison of the annotated kink positions

Distance	HELANAL (270 helices)	PDS (268 helices)
0 AA	84 (31%)	51 (19%)
1 AA	90 (33%)	106 (40%)
2 AA	48 (18%)	56 (21%)
3 AA	15 (6%)	20 (8%)
4 AA	11 (4%)	17 (6%)
>4 AA	22 (8%)	18 (6%)

Absolute distance in amino acids (AA) to our manually created data set for all helices labeled automatically as kinked. The most (>90%) are within 1 helical turn.

5.2.3 *Application of SVMs*

Table 5.4 gives an overview of the results for the nine best string kernels for SVMs and the neural network of TMKink. Annotating the extracted helices in our data set manually yields a much higher balanced accuracy and F-score compared to both automated methods, HELANAL and PCA. Applying our method to the data set

from the MC-HELAN Web site, we achieve the same sensitivity but a much lower specificity and, hence, a lower balanced accuracy – even compared to HELANAL. The higher F-score is mainly due to the larger number of kinked helices in their data set. This supports our assumption that automated kink detection from structural information is still too error prone and noisy to be useful for training statistical classifiers.

Table 5.4: Prediction results for the four data sets

Rank	MDS	HELANAL	PDS	MC-HELAN
1	A_G0_1 0.820 (0.801)	K4_M2_W0 0.766 (0.707)	K4_M2_W0 0.630 (0.635)	K4_M2_W0 0.742 (0.811)
2	K4_M2_W0 0.811 (0.791)	A_G0_05 0.759 (0.699)	A_G0_1 0.616 (0.604)	A_G0_1 0.719 (0.802)
3	A_G0_05 0.808 (0.788)	A_G0_1 0.755 (0.693)	K4_M2_W1 0.606 (0.619)	K3_M1_W0 0.710 (0.779)
4	A_G0_01 0.786 (0.770)	A_G0_01 0.751 (0.692)	K5_M2_W1 0.605 (0.548)	K3_M1_W1 0.690 (0.765)
5	K4_M2_W1 0.778 (0.748)	K1234_M0_W1 0.742 (0.679)	K123_M0_W0 0.604 (0.606)	K1234_M0_W1 0.690 (0.786)
6	K12345_M0_W1 0.767 (0.734)	K4_M2_W1 0.737 (0.669)	K3_M1_W0 0.601 (0.611)	K4_M2_W1 0.688 (0.770)
7	K1234_M0_W1 0.765 (0.731)	K12345_M0_W1 0.734 (0.668)	K1234_M0_W1 0.600 (0.542)	K12345_M0_W0 0.686 (0.792)
8	K3_M1_W0 0.752 (0.730)	K3_M1_W0 0.733 (0.667)	K12345_M0_W0 0.600 (0.578)	K1234_M0_W0 0.679 (0.779)
9	K5_M2_W1 0.748 (0.703)	K123_M0_W1 0.733 (0.667)	K5_M2_W0 0.597 (0.529)	A_G0_05 0.673 (0.753)
-	TMKink 0.714 (0.707)	TMKink 0.691 (0.641)	TMKink 0.618 (0.621)	TMKink 0.630 (0.724)

Kernel name as well as the corresponding balanced accuracy and Fscore (in brackets) are given. The colors are due to the balanced prediction accuracy from yellow (low) to green (high). The last row shows the results of the TMKink method.

Interestingly, the F-score we achieve significantly exceeds the one reported by Langelaan et al. and demonstrates the usefulness of string kernel-based SVMs, even more with respect to the neural networks results of TMKink, where the balanced accuracy and F-score decreases dramatically. But again, MDS performs best, while the data set of Langelaan and co-workers has the highest F-score.

The three alignment kernels (with different parameter γ) are ranked in the top four. This seems to indicate that the order of the amino acids might play a decisive role, information that is lost when using k-mer kernels. Whereas a combination of mismatch and spectrum kernel slightly decreases the performance in most cases, we were able to improve the results remarkably using a combined spectrum kernel with different k-mer sizes. For example, K12345_M0_W1 (76%) is better than the nonweighted version K12345_M0_W0 (73%) and compared to the best single spectrum kernel K3_M0_W1 (62%). Furthermore, mismatches yield better results than weighting the occurrences: especially using k-mers of size 4 and allowing 2 mis-

matches without weighting the occurrences (K4_M2_W0) seems to be a good choice for kink prediction.

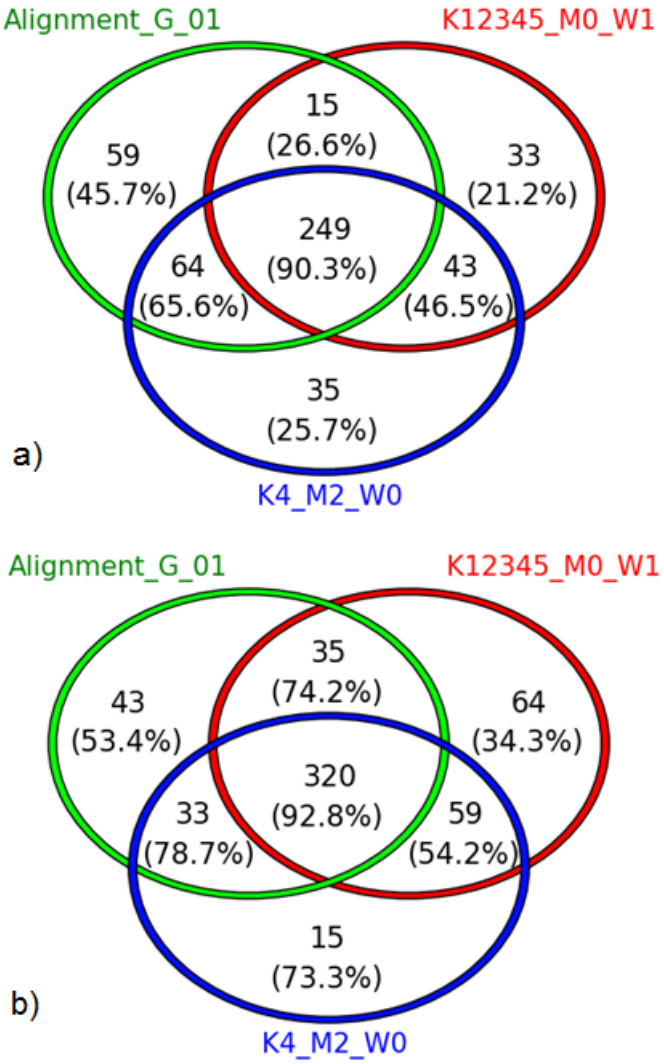


Figure 5.6: Venn diagram of the prediction results for kinked (a) and nonkinked (b) helices. The number in brackets is the percentage rate of correctly identified kinked (nonkinked) helices.

Figure 5.6 gives a Venn diagram of the prediction results for the best alignment, mismatch, and spectrum kernel. About 90% of the helices predicted as kinked by all three string kernels are indeed kinked. For canonical helices, this number is even higher. Compared to the number of kinked and nonkinked helices in MDS, nonkinked helices are predicted with a balanced accuracy of 63.0% (64.4%) by all three kernel methods. Taking the majority vote of these three kernels, we can predict nonkinked helices with a balanced accuracy of 82.1% (82.6%). These results strongly indicate that string kernel-based SVMs yield a very stable and adequate method for kink prediction.

Table 5.5: Confusion Matrix of MDS for Proline and Nonproline Kinks

kernel function	TP	FN	TN	FP	sensitivity	specificity	balanced accuracy
K4_M2_W0 proline	192	14	19	18	93.2	51.4	72.3
Alignment_G0_1 proline	181	25	33	4	87.9	89.2	88.5
K4_M2_W0 nonproline	104	47	347	77	68.9	81.8	75.4
Alignment_G0_1 nonproline	117	34	339	85	77.5	80	78.7

Another question of interest in kink prediction is the sensitivity and specificity for proline and nonproline kinks. As mentioned above, we defined a proline-induced kink if proline occurs in the range of ± 5 residues away from a specified kink position. Table 5.5 shows that the alignment kernel is significantly better than the k-mer kernels in predicting proline kinks due to the very high specificity. We suppose the exact position of proline to play an important role, which is confirmed by the neighborhood analysis (Fig. 5.3), where proline occurs mainly 0-5 residues after a kink. Nonproline kinks have been detected with a high and balanced sensitivity and specificity. These results are very promising, although we are not focusing on the exact kink position in this work. In particular, our approach reveals SVMs to be capable to find other general features besides the occurrence of proline in the sequence.

5.2.4 *Kink Neighborhood*

Today, the main influences for kink formation are still unknown. In fact, it is even unclear whether kinks are a very global (influences over different helices and nonhelical parts), a mostly local (influences only inside the same helix), or a very localized (influences only from a few residues around the kink) effect. To decide whether a few residues around the kink are enough to classify into kinked and nonkinked, we created further data sets containing only the so-called core subsequence (CDSX) of each helix, where X denotes the length of the subsequences. In cases of kinked helices, the kinked residue corresponds to the center of the considered subsequence. For nonkinked helices, we decided to set the center of the complete helix to the center of the subsequence, because this part is mostly in the membrane center and usually more important and reliable for the stability of the helical structure than regions at the end of a helix. Compared to our first results, we obtain a lower balanced accuracy and F-score but in many cases still over 75% with a maximal F-score of 0.74 (see Table 5.6). These results indicate that a large part, but not all, of the effects behind kink formation seems to be very localized. In addition, we classified each subsequence by TMKink and got clearly lower prediction results. Because Bowie and co-workers predicted kinks in a range of ± 4 residues and also Langelaan et al. reported their results with this window size, we suppose that their results are comparable to the one of CDS9.
It is noticeable that only in one case, the alignment kernel achieves the 10th best result. Hence, we suppose this kernel to be influenced by the length of a sequence. CDS9-CDS13 work very similar, which means that we have to consider 4-6 amino

Table 5.6: Prediction results for CDSX

Rank	CDS7	CDS9	CDS11	CDS13	CDS15
1	K4_M2_W0 0.729 (0.694)	K1234_M0_W1 0.771 (0.737)	K4_M2_W0 0.769 (0.734)	K4_M2_W1 0.779 (0.740)	K4_M2_W1 0.750 (0.702)
2	K4_M2_W1 0.725 (0.688)	K5_M2_W0 0.767 (0.729)	K4_M2_W1 0.767 (0.730)	K4_M2_W0 0.770 (0.731)	K4_M2_W0 0.748 (0.702)
3	K12345_M0_W0 0.724 (0.687)	K123_M0_W1 0.765 (0.732)	K1234_M0_W0 0.766 (0.727)	K1234_M0_W0 0.765 (0.725)	K5_M2_W0 0.745 (0.691)
4	K1234_M0_W1 0.723 (0.688)	K4_M2_W0 0.760 (0.724)	K12345_M0_W0 0.766 (0.726)	K123_M0_W1 0.764 (0.726)	K12345_M0_W1 0.743 (0.696)
5	K123_M0_W1 0.721 (0.685)	K12345_M0_W0 0.759 (0.718)	K12345_M0_W1 0.759 (0.721)	K5_M2_W0 0.761 (0.714)	K1234_M0_W1 0.742 (0.695)
6	K123_M0_W0 0.721 (0.686)	K4_M2_W1 0.756 (0.724)	K123_M0_W1 0.757 (0.723)	K12345_M0_W0 0.757 (0.714)	K12345_M0_W0 0.737 (0.682)
7	K12345_M0_W1 0.720 (0.685)	K12345_M0_W1 0.756 (0.721)	K1234_M0_W1 0.753 (0.716)	K5_M2_W1 0.753 (0.699)	K123_M0_W1 0.737 (0.691)
8	K3_M1_W0 0.720 (0.687)	K1234_M0_W0 0.755 (0.715)	K5_M2_W0 0.752 (0.709)	A_G0_1 0.751 (0.712)	K123_M0_W0 0.736 (0.692)
9	K1234_M0_W0 0.712 (0.673)	K5_M2_W1 0.745 (0.699)	K3_M1_W0 0.748 (0.713)	K1234_M0_W1 0.743 (0.692)	K5_M2_W1 0.732 (0.669)
-	TMKink 0.0 (0.0)	TMKink 0.648 (0.508)	TMKink 0.701 (0.638)	TMKink 0.709 (0.672)	TMKink 0.672 (0.666)

Kernel name as well as the corresponding balanced accuracy and Fscore (in brackets) are given. The colors are due to the balanced prediction accuracy from yellow (low) to green (high). The last row shows the results of the TMKink method.

Table 5.7: Confusion Matrix of CDS11 for Proline and Nonproline Kinks

kernel function	TP	FN	TN	FP	sensitivity	specificity	balanced accuracy
K4_M2_W0 proline	188	9	23	20	95.4	53.5	74.5
AlignmentG0_1 proline	182	15	31	12	92.4	72.1	82.2
K4_M2_W0 nonproline	60	80	344	70	42.9	83.1	63.0
AlignmentG0_1 nonproline	61	79	306	108	43.6	73.9	58.7

acids to the left and to the right of a specified kink. This observation correlates with the kink environment plot in Figure 5.3. Kernel K4_M2_W0 is again a good choice. The resulting confusion matrix for CDS11 for proline and nonproline kinks is illustrated in Table 5.7. While proline kinks using the mismatch string kernel K4_M2_W0 are predicted with a slightly higher balanced accuracy, we found a decrease in all other cases. The reason is, in particular, the very low sensitivity for nonproline kinks indicating that nonproline kinks are not local ones and that a larger sequence range has to be taken into account.
To confirm this supposition, we trained a statistical model on a data set containing only the nonproline sequences of MDS and CDS11 (see Table 5.8). These data sets are called NP_MDS and NP_CDS11, respectively. Applying all kernels, we obtained a significantly better performance including the whole helical sequence. In this case, more than 55% of nonproline kinks were predicted correctly. Focusing on just 5

Table 5.8: Confusion Matrix of NP_MDS and NP_CDS11

kernel function	TP	FN	TN	FP	sensitivity	specificity	balanced accuracy
K4_M2_W0 NP_MDS	67	54	316	41	55.4	88.5	71.9
AlignmentG0_1 NP_MDS	73	48	328	29	60.3	91.9	76.1
K4_M2_W0 NP_CDS11	45	68	315	38	39.8	89.2	64.5
AlignmentG0_1 NP_CDS11	43	70	281	72	38.1	79.6	58.8

neighboring residues this value decreased to just below 40%. The lower sensitivity and higher specificity compared to the usage of the complete data sets might be a result of the extremely unbalanced data set.
In addition, we compared the helix length of proline and nonproline kinked helices, finding an average length of 26.4 and 27.1, which results in a nonsignificant difference (p-value: 0.15). Hence, the helix length itself does not influence the result, implying that for nonproline kinks, the whole sequence seems to be very important for a better prediction.

5.2.5 *Detecting the Exact Kink Position*

In principle, SVMs can also be used for detecting the exact kink position in the helical sequence in addition to the binary kink/nonkink classification. The last results show that an SVM is able to predict also smaller sequences with a high balanced accuracy correctly. Focusing on the results of CDS9, we can predict kinks in a range of ± 4 residues with a balanced accuracy of more than 75%, which is higher than reported accuracies in former studies. However, in further studies we tried to be even more precise. Our idea was to use a window of a specific size to create all subsequences of a helix while iterating over it. The subsequences will be labeled as kinked if and only if it contains the kinked residue. After applying SVMs to this modified data set, we retranslated the prediction result to the whole helix. Unfortunately, we got about the same result as before since there is still a large degree of noise due to the very short sequences, which have in some cases different labels in various helices. We assume that we reached the limit of predicting kinks from sequence. One has to keep in mind, as stated already above, that also inter-helical interactions play most probably a crucial role in kink formation.

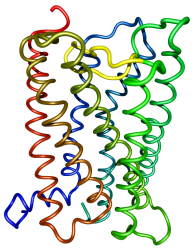
5.3 CONCLUSION

Our work gives new insights on kinks in α-helical membrane proteins. First, our data set analysis affirms the great importance of proline for distortions but reveals also a disproportionately high occurrence of glycine and serine. Moreover, the data set analysis can help to assess and improve homology models by incorporating the gained information. Some of these results are already confirmed by the findings of the related work of Langelaan et al, e. g., the high occurrence of proline a few positions after a kink.

Furthermore, we have developed and validated a new kink prediction method using string kernels for SVM and our manually annotated data set. The very high consensus of all applied string kernels demonstrates that there is much information about kinks coded in the amino acid sequence of a helix. Most importantly, using string kernels allows us to detect also nonproline kinks with a high accuracy, where the most of the previously published methods more or less failed. The basis of these considerably improved results is our manually created data set and the usage of string kernels, which is demonstrated in the comparison between both manually and automatically annotated data sets as well as different methods. Nevertheless, we agree with Langelaan and co-workers that the helical sequence is only one factor and that tertiary interactions or the spatial environment (membrane) cannot be neglected, which is confirmed by the SASA analysis of the different helix types.
Finally, we provide a large data set for further studies. This, for example, can be used to develop and evaluate future algorithms for determining kinks from three-dimensional structures automatically.

In the following chapter, we will present a new approach to model kinks during the rearrangement of the helices.

6 *Fragmental GPCR modeling*



In this chapter, we present our new fast and fully automated approach to model G-protein coupled receptors. More importantly, it replaces the optimization of rigid helices by a more sophisticated method that allows helices to change their conformation during this process. Thus, helices that are wrongly kinked, e. g., helices obtained by molecular dynamics simulation, have the possibility to move to the right position based on their inter-helical contacts. We will demonstrate that the models we obtain are better than all previously published ones and that the optimization method we developed does not need any interaction by the user. Next, we will describe our so-called *fragmental GPCR modeling approach* step by step and present and discuss the results we obtained applying this new algorithm.

The idea of our approach is to be completely independent of any prediction method for initial models. Good starting structures are essential for all optimization methods since we cannot compute and assess all possible conformations. Therefore, we want to have again a closer look at all known GPCR crystal structure to possibly deduce some general structural features. Fortunately, we have at least one representative from three of the four rhodopsin-like subfamilies. Five structures (PDB IDs: 1U19, 2RH1, 2VT4, 3PBL, 3RZE) belong to the α-, one structure (PDB ID: 3EML) to the β- and one structure (PDB ID: 3ODU) to the γ-subfamily. Structural features found in all of these structures should for sure hold for the other proteins, which belong to these subfamilies and most probably also for those of the δ-subfamily.

6.1 COMPUTATION OF INITIAL MODELS

To generate appropriate starting conformations, we did not rely on any prediction method, e. g., for the secondary structure or the orientation of the helix, because they fail in many cases as demonstrated in Chapter 4.
The results of our comparison studies of GPCR crystal structures have already shown high structural similarity of the overall fold (see Figure 4.3) as well as the helical ends (see Table 4.3).
Focussing on the conformation of the single helices, it is obvious that H1 has the most diverse kink in the outer membrane side as illustrated in Figure 6.1. This is mainly due to insufficient stabilization factors during crystallization in this region, which can lead in the worst case to an unrealistic and excessive kink as stated by Warne et al. for the case of 2VT4.[10] Hence, this difference in H1 is most probably not as distinctive in nature as one might deduce from the available crystal structures. Here, it is important to keep in mind that crystal structures are also a kind of model, in particular, because proteins are always in motion.

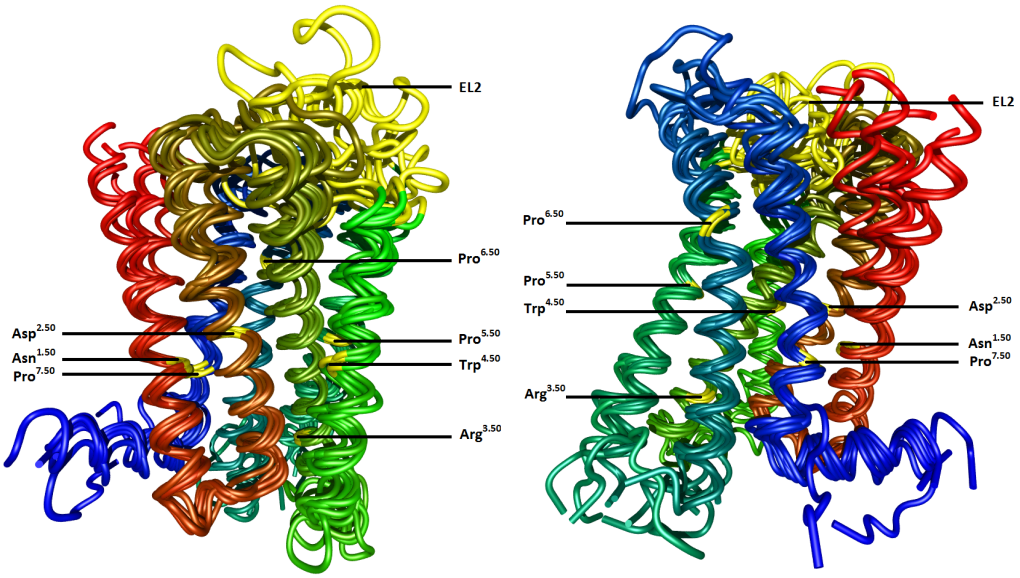


Figure 6.1: The GPCRs are mapped by the C_α-atoms of the seven most conserved residues X.50 (marked yellow). The most diverse region is the extracellular loop 2 (ECL2), which is also highlighted.

As illustrated on the left side of Figure 6.1, H4 is twisted differently in the seven GPCRs, leading to a high RMSD value when comparing these. Although it has two potential binding pocket residues denoted by Rognan (see Figure 4.2), it is not always involved in the ligand binding mode as, for example, in retinal bound bovine rhodopsin. Because the influence of H4 is too strong in the final evaluation of the model's quality, one might prefer to exclude H4 as done by Shacham et al., who reduced the C_α-RMSD value in this case from 3.87Å to 3.2Å.[32] However, since we focus on fully automatic GPCR modeling an exclusion of a whole helix is too strong as a restriction. Besides the C_α-RMSD values of the complete model (excluding the loop regions), we decided to compute the C_α-RMSD of the binding pocket residues as well as the RMSD value of the binding pocket residues where only the hydrogen atoms were excluded in the following analysis.
Apart from H4, all other helices are only slightly differently kinked and/or twisted in the outer membrane side, while they fit very well in the intracellular (IC) side. In contrast to other approaches, where canonical helices[32] or simulated helices[164] were used in the beginning of the GPCR modeling procedure, our idea for obtaining suitable initial conformations is to take advantage of this structural similarity.

Therefore, we used *Artificial Evolution* (AE), which is a homology modeling algorithm that takes a template structure as well as an alignment of the template and target sequence as input. Like in evolution, the template structure is then converted into the target structure in a step-by-step manner. In each step, all remaining point mutations (including deletions and insertions) of amino acids are performed and assessed. The mutation causing the smallest change in the fitness score (ΔE) is then chosen. Consequently, this procedure takes into account that evolution happens

most likely in very small steps. The process is finished as soon as all mutations have been applied, i. e., a structure of the target is achieved.
In his Master's thesis, Baldauf showed that the best results can be achieved if insertions and deletions are not taken into account, which makes this approach even easier.[165] In this case, it is not necessary to manually refine the alignment of both sequences since the bijection of residues of the template and target helix is fixed by the most conserved residue X.50 according to the Ballesteros-Weinstein nomenclature. Figure 6.1 illustrates the high structural fit of the conserved residues (highlighted in yellow). Moreover, Baldauf demonstrated that a backbone optimization after each mutation does not improve the results significantly, which emphasizes again the high similarity between the helices in the GPCR crystal structures.

Table 6.1 gives all RMSD values of our initial models for bovine rhodopsin (PDB ID: 1U19) and the human CXCR4 chemokine receptor (PDB ID: 3ODU) after the application of the simplified artificial evolution algorithm to the six other template structures and using SCWRL4[166] to rearrange the side chains after the AE step.

Table 6.1: RMSD values obtained by artificial evolution

1U19	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
2RH1	1.28	0.92	0.92	1.15	1.00	1.24	1.73	2.28	2.15	2.98
2VT4	1.42	0.98	0.95	1.34	1.00	1.19	1.42	2.37	2.19	2.76
3EML	1.63	1.85	1.53	1.62	1.61	1.85	0.99	2.54	2.47	3.18
3ODU	1.31	2.51	1.23	2.65	1.07	0.99	1.33	2.42	2.55	3.28
3PBL	0.84	0.87	0.79	0.77	0.84	0.94	1.08	1.75	1.73	2.61
3RZE	1.38	1.17	1.15	1.88	1.09	0.94	1.40	2.03	1.96	2.48

3ODU	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.44	2.36	1.23	3.25	1.18	1.03	1.23	2.70	2.65	3.27
2RH1	2.13	2.26	0.61	2.90	0.83	1.29	1.53	3.17	2.88	3.62
2VT4	2.20	2.14	0.75	3.06	0.76	1.49	1.47	3.37	2.89	3.59
3EML	2.09	2.17	1.72	3.05	2.19	1.63	1.75	3.29	2.89	3.60
3PBL	1.59	1.96	0.79	3.15	0.69	1.42	1.55	2.67	2.23	2.71
3RZE	2.26	2.27	0.71	3.27	0.86	1.04	1.85	2.87	2.40	3.19

The C_α-RMSD of each helix and the whole model (PDB ID: 1U19 and 3ODU), when using artificial evolution. In addition, we give the C_α-RMSD and the RMSD of the all heavy atoms of the binding pocket residues defined by Rognan. The side chains were optimized using SCWRL4.[166]

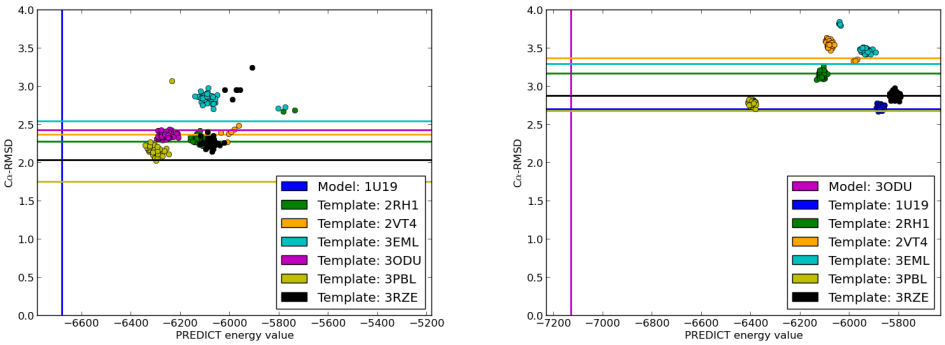
Since the templates are evolutionarily more closely related to bovine rhodopsin than to the human CXCR4 chemokine receptor, the corresponding C_α-RMSD values are, as expected, smaller on average (2.23Å compared to 3.02Å, respectively). But it is very surprising that the human dopamine D3 receptor (PDB ID: 3PBL) is in both cases the best template structure, leading to an C_α-RMSD value of 1.75Å for bovine rhodopsin, which is significantly smaller than any published in silico model we know of. Except for one model, the C_α-RMSD value is decreased when

focussing on the binding pocket residues only, whereas this decrease is on average stronger for the human CXCR4 chemokine receptor models (-0.30Å) compared to the bovine rhodopsin models (-0.08Å). This is due to the high C_α-RMSD value of H4 in 3ODU (almost always at least 3.0Å), which has only 2 putative binding pocket residues as mentioned before. The RMSD values of the binding pocket residues defined by Rognan et al. is indeed higher but since almost all docking algorithms take receptor side chain flexibility into account these values are only given to get a first impression of the binding pockets' reasonability. Unfortunately, none of the other studies gives RMSD values of the binding pocket residues such that we are not able to exactly assess our results.

These already reasonable results are now used as input for our optimization procedure. Although we expect to improve these models, the simple energy function we use might not be able to distinguish between the models at this level of detail since the application on the crystal structures conformation already showed that models with lower energy values but higher C_α-RMSD values (between 1.0Å and 2.0Å) can be generated. Nevertheless, a difference in performance of the rigid and fragmental modeling approach should be recognizable. Moreover, we run two different tests here: one with side chain optimization (SCO) and one without. The idea is that our initial models are already in an appropriate conformation such that the side chain positions computed by SCWRL4 might be more meaningful than those rearranged during the simulated annealing optimization, where the side chains exist only in a reduced representation.

Before describing our new algorithm, which is able to handle kinks during the optimization process, we show the results obtained by applying our simulated annealing algorithm, where helices were treated as one rigid body. On the one hand, we want to check if the PREDICT scoring function is able to distinguish between these initial models and, on the other hand, we want to have a closer look on 3PBL, which is in both cases the template with the lowest C_α-RMSD value. Because we do not want to dump all the available data to the reader, we only present the results obtained using the SCWRL side chain positions in these test cases. We discuss the side chain placement in comparison when applying our new optimization algorithm in Section 6.3.1.

We achieved diverse results (see Figure 6.2) when applying the single rigid body simulated annealing optimization procedure. Depending on the starting conformation, almost all final structures are very close to each other. The reason for this effect is both the already proper conformation of the helices and the exclusion of SCO. Thus, the helices have no impulse (with regard to the PREDICT scoring function) to change the overall fold significantly, when treating them as rigid cylinders. The C_α-RMSD values increase on average, in particular when using 3PBL as template to model 1U19. However, the template 3PBL has in both test cases the lowest energy values and almost the lowest obtained C_α-RMSD value. Only when using template 1U19 to model 3ODU we generate models with a non-significant lower C_α-RMSD value. Whereas the PREDICT energy function completely failed in arranging canonical helices as it was done in the PREDICT procedure (see Figure 4.12 and 4.14), it



(a) 1U19 **(b)** 3ODU

Figure 6.2: 50 restrained single rigid body SA runs using the side chain positions computed by SCWRL4 applied to the initial conformation created by AE. The horizontal lines represent the C_α-RMSD values of the corresponding initial conformation. The vertical line is the energy value of the crystal structure.

works quite well to assess our initial models with the SCWRL4 side chain positions. The only exception is 3EML (see Figure A.3), where we achieve the lowest energy values for models created from the template 3ODU. Unfortunately, these models have a high C_α-RMSD values. Obviously, there are some interactions that are wrongly assessed by the PREDICT energy function.

Summarizing these discoveries, it is quite obvious that we can build promising initial models using preprocessed helices, i. e., helices, which are already in a proper conformation. Instead of helices obtained by MD simulation, we used the conformation of helices from other published GPCR crystal structures. This method is very fast and straightforward and can achieve even better results when more crystal structures are available or when helices of different crystal structures are combined. For example, when modeling 1U19, we have for each helix at least one template structure that has a C_α-RMSD smaller than 1Å. Most important is, however, that we use the already available folds to map the helices to. Hence, we are independent of any prediction methods for helix length or their orientation. Of course, this is not ab initio modeling in the true sense but in their current state pure ab initio prediction methods introduce more errors than they help to improve the quality of initial models. On the other hand, it is also no pure homology modeling approach since we do not rely on sequence alignments, which is said to be one of the most important steps in homology modeling.

Next, we focus on our new fragmental GPCR modeling approach. The idea of this approach is to optimize the obtained initial models with respect to putative kinks. As already described above (see also Figure 6.1), GPCRs overlay very well in the IC side, while they are slightly differently kinked in the outer membrane side. Because ligands bind mainly to the latter area, it is important to model this region well.

Hence, to represent a helix as only one monolithic rigid body is not sufficient to model these differences in an appropriate manner.

6.2 MATHEMATICAL BACKGROUND

In this section we describe the modeling of cylinder fragments. As stated above, all fragments have three degrees of freedom for rotation, whereas only the first one has three additional degrees of freedom for translation. A rotation around the x-, y- or z-axis can be performed by multiplying with the following rotation matrices.

$$R_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

$$R_y(\beta) = \begin{pmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{pmatrix}$$

$$R_z(\gamma) = \begin{pmatrix} \cos(\gamma) & -\sin(\gamma) & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

To describe a rotation in an orthogonal coordinate system we need a triplet of angles and axes. There exists many different conventions for the definition of rotation axes and rotation order. One of the most common ones is described by Rose, where all rotations are performed in a counter-clockwise fashion around fixed axes.[167] We rotate only in a very small range of about $\pm 15°$, and hence avoid the so-called *Gimbal Lock* problem, i. e., the loss of a degree of freedom during rotation. Moreover, since we rotate and translate all atoms in each step, the usage of quaternions has no distinct advantage anymore.[168] The stability in the computation of new coordinates is guaranteed in our implementation due to the application of the whole rotation on the initial coordinates in each step.
Let $p = (p_x, p_y, p_z)$ be a point in 3D, then its new coordinates $p' = (p'_x, p'_y, p'_z)$ are computed according to Rose's definition as follows:

$$\begin{pmatrix} p'_x \\ p'_y \\ p'_z \end{pmatrix} = R_z(\phi)R_y(\theta)R_z(\psi) \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{19}$$

where ψ, θ, and ϕ are the corresponding rotation angles and $t = (t_x, t_y, t_z)$ is the translation vector. The rotation angles ψ and ϕ are defined in the range $(0, 2\pi)$, whereas θ is defined in the range $(0, \pi)$. In the case where we represent a helix by a single cylinder - as it was done in PREDICT and also in this thesis until now -, this is everything we need for the simulation.

However, we want to allow helices to change their conformation during the simulation, in particular, in the outer membrane side. Therefore, we split the helix into predefined fragments (see Figure 6.3), where the first one (colored blue) is always the one in the IC side. For this fragment, we computed the geometric center of the backbone atoms and shifted it accordingly to the point of origin. Afterwards, it is rotated such that the helical axis, again computed by Principal Component Analysis (PCA),[148] is mapped onto the z-axis as illustrated in Figure 6.4a. All subsequent fragments, in our example fragment 2 (red) and 3 (green), are just translated such that their first atom is placed to the point of origin (Figure 6.4b and 6.4c). By this procedure, these atoms serve as anchor points as they will not change their position when applying rotation matrices to these fragments. The helix fragments in the current orientation are our starting points for all further computations. Since we translated each fragment once, they already have a translation vector different from zero. The rotation angles of the first fragment are also non-zero due to the already applied rotations, while the rotation angles of all following fragments are 0.

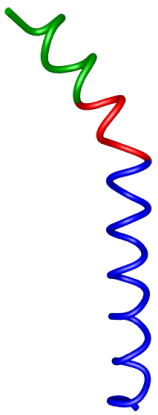
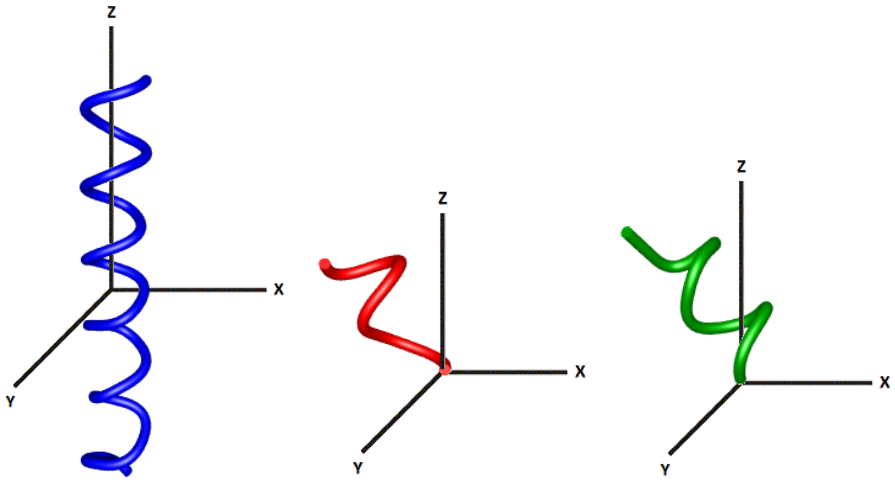


Figure 6.3: Exemplary helix consisting of three fragments.



(a) First fragment **(b)** Second fragment **(c)** Third fragment

Figure 6.4: The three fragments of a split helix. The helix axis of the first is oriented along the z-axis. The first atom in each of the following fragments is moved into the coordinate origin.

When the values of the degrees of freedom are changed in the simulation process, we compute the new position of each atom in the following way: We start rotating the last fragment according to its new values and directly connect it to the previ-

ous one by applying the translation vector of its anchor atom. The rotation matrices belonging to the previous fragment are then applied to the already combined fragment. This procedure is repeated until the first fragment is finally translated. The new coordinates p' of an atom with initial position p belonging to the second fragment are therefore computed as follows:

$$\begin{pmatrix} p'_x \\ p'_y \\ p'_z \end{pmatrix} = \underbrace{R^1_z R^1_y R^1_z}_{\substack{\text{rot. mat.} \\ \text{first fragm.}}} \left[\underbrace{R^2_z R^2_y R^2_z}_{\substack{\text{rot. mat.} \\ \text{second fragm.}}} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} + \underbrace{\begin{pmatrix} t^2_x \\ t^2_y \\ t^2_z \end{pmatrix}}_{\substack{\text{fixed by} \\ \text{anchor atom}}} \right] + \underbrace{\begin{pmatrix} t^1_x \\ t^1_y \\ t^1_z \end{pmatrix}}_{\substack{\text{trans. vec.} \\ \text{first fragm.}}} \quad (20)$$

The hinges were optimized in each simulated annealing step in the same way as described for the rotation angles of the first fragment. Deviations from the starting angle for each hinge have also been penalized through the restraint function given in Equation 17. However, the factor c_f in this equation was reduced to 5 since we explicitely want to allow helices to change their conformation in the outer membrane side.

Before we present the results of our new algorithm, we have to discuss one drawback of this method - the placement of proper hinges. At the time of writing of this thesis, the prediction method we developed (see Chapter 5) achieves the best results in predicting kinks from protein sequence. Nevertheless, even this method has a prediction accuracy of only about 80% and is still not able to annotate the exact kink position with a high accuracy. Moreover, it is most probably not sufficient to place only one kink per helix. For example, if a helix of the template is kinked at position X but the target is distorted at position Y, we need at least two hinges to reduce the kink at position X and to introduce one at position Y.

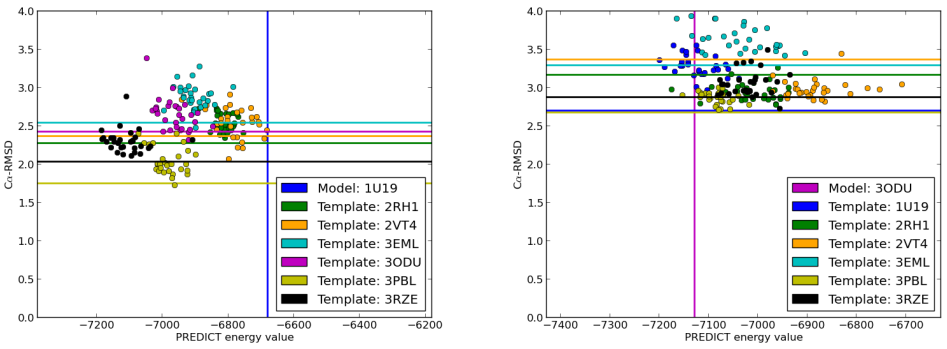
Hence, the question arises how many hinges per helix we need to model the binding pocket in a flexible way without too many additional degrees of freedom. The first fragment, the one in the IC side, was set to be at least 15 residues long since this region is structurally conserved and is not in our focus. On the other hand, the last fragment should have a size of at least one turn (4 residues). Hence, the hinges are placed in a range of about 10-15 residues. Setting 3-5 hinges per helix, one can expect to have one hinge per helical turn on average. Due to this large amount of hinges, it is most probably not important where exactly we set a hinge. However, to be able to reproduce our results and to analyze if the rigid approach can be improved by hinges, we tried to set the hinges for our test example at specific positions.
Therefore, we compared the target and template structure as follows: First, we reduced their representation according to Herzyk et al.[36] We then mapped for each four consecutive virtual C_α-atoms the first three of the template ($\vec{C}^{Te}_1$, $\vec{C}^{Te}_2$, $\vec{C}^{Te}_3$) onto the corresponding target atoms ($\vec{C}^{Ta}_1$, $\vec{C}^{Ta}_2$, $\vec{C}^{Ta}_3$). Afterwards, we calculated

the angle between the two vectors $\vec{V}_1 := \vec{C}_4^{Te} - \vec{C}_3^{Te}$ and $\vec{V}_2 := \vec{C}_4^{Ta} - \vec{C}_3^{Ta}$ and set the hinge at the $\vec{C}_3^{Ta}$ if the angle $\measuredangle(\vec{V}_1, \vec{V}_2)$ was larger than 20°. Since helix 3 is very similar throughout the whole GPCR family, we used only one hinge to optimize this helix. This hinge was set to the C_α-atom with the largest obtained angle even when the angle was smaller than 20°. The other helices were equipped with up to 5 hinges, which were allowed to be at neighboring residue positions.

6.3 RESULTS

In the following two sections, we present and discuss the results obtained by applying our new optimization method to the initial structures with both options, including side chain optimization and using the side chain positions computed by SCWRL4.

6.3.1 *Side chain optimization*



(a) 1U19 **(b)** 3ODU

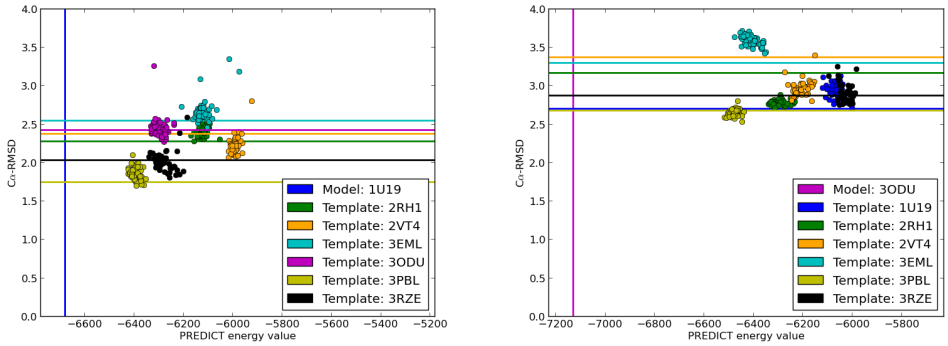
Figure 6.5: 50 restrained fragmental modeling SA runs including SCO applied to the initial conformation created by AE. The horizontal lines represent the C_α-RMSD values of the corresponding initial conformation. The vertical line is the energy value of the crystal structure.

As illustrated in Figure 6.5, the final models are, as expected, more diverse even if they have been generated from the same initial conformation. This led to a partially substantial shift of the template structures such that 3RZE is now the template with the smallest energy value in average when modeling 1U19. However, 3RZE is only the second best template with regard to the RMSD values. In case of 3ODU, it is even more problematic since the RMSD values are not correlated with their energy values at all. Another effect of using SCO is that the energy values of the final models are often lower than the ones of the crystal structure. These results emphasize again that the energy function is useful for short post-optimization, but already when allowing rearranging side chains, it fails. Due to computation time it is not possible to optimize the side chain positions during the simulated annealing

procedure using SCWRL4, and hence the question arises if it is sufficient to use the SCWRL4 side chain positions computed initially. Moreover, when we exclude SCO we limit the conformational space for optimization drastically, and it is debatable, whether we can see an effect of our new modeling algorithm compared to the single rigid body optimization.

6.3.2 *SCWRL side chain positions*

The results of our new approach without side chain optimization for bovine rhodopsin (PDB ID: 1U19) and the human CXCR4 chemokine receptor (PDB ID: 3ODU) are presented in Figure 6.6 and will be compared to those in Figure 6.2 in the following.



(a) 1U19 **(b)** 3ODU

Figure 6.6: 50 restrained fragmental modeling SA runs using the side chain positions computed by SCWRL4 applied to the initial conformation created by AE. The horizontal lines represent the C_α-RMSD values of the corresponding initial conformation. The vertical line is the energy value of the crystal structure.

Again, we obtain the lowest PREDICT energy values using the template 3PBL, but with our new approach it correlates even better with the C_α-RMSD values, i. e., the models obtained with this template are also the closest to their corresponding crystal structure. The models created from the same template are more diverse due to the higher number of degrees of freedom. However, they often converge from different starting conformations to a similar final model, for example, the models based on the templates 2VT4 and 3RZE when modeling both 1U19 or 3ODU. Another nice (although not significant) effect of our new approach is that the C_α-RMSD values are more often smaller than that of the starting conformations, which was very uncommon in the old approach. This shows the importance of additional flexibility in the binding pocket region.

The successful application is further emphasized when analyzing the models of 3EML (see Figure A.4). In the rigid approach the models with the highest C_α-RMSD are assessed best, whereas the models with the lowest PREDICT energy value ob-

tained using our fragmental modeling approach have a much lower C_α-RMSD value (decrease of more than 0.4Å). In this case, we clearly notice the difference in the quality of the two approaches.

To compare the results in more detail, we first have a short look at Tables 6.2 and 6.3. On the one side, the C_α-RMSD values for the obtained models are better on average when using our new approach (in 8 of 12 cases), but, on the other side, this improvement is neither reflected in the C_α-RMSD of the binding pocket residues (6 of 12) nor in their side chain positions (4 of 12). Since we are explicitely interested in this region, these results are disappointing at first glance. However, we are not interested in all of these models because we are able to exclude most of the inappropriate ones due to their high energy values. To do so, we additionally applied our gradient-based optimization procedure and selected for each crystal structure only the model with the lowest PREDICT energy value for both the rigid helix approach and our new fragmental modeling approach. Our final results are given in the Table 6.4.

Table 6.2: RMSD values of the rigid GPCR approach

1U19	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
2RH1	1.28	0.92	0.92	1.15	1.00	1.24	1.73	2.28	2.25	2.99
2VT4	1.42	0.98	0.95	1.34	1.00	1.19	1.42	2.32	2.13	2.71
3EML	1.63	1.85	1.53	1.62	1.61	1.85	0.99	2.83	2.95	3.54
3ODU	1.31	2.51	1.23	2.65	1.07	0.99	1.33	2.36	2.34	3.18
3PBL	0.84	0.87	0.79	0.77	0.84	0.94	1.08	2.18	2.16	2.90
3RZE	1.38	1.17	1.15	1.88	1.09	0.94	1.40	2.25	2.01	2.51

3ODU	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.44	2.36	1.23	3.25	1.18	1.03	1.23	2.72	2.60	3.22
2RH1	2.13	2.26	0.61	2.90	0.83	1.29	1.53	3.15	2.88	3.58
2VT4	2.20	2.14	0.75	3.06	0.76	1.49	1.47	3.54	3.06	3.73
3EML	2.09	2.17	1.72	3.05	2.19	1.63	1.75	3.83	3.29	4.02
3PBL	1.59	1.96	0.79	3.15	0.69	1.42	1.55	2.78	2.31	2.81
3RZE	2.26	2.27	0.71	3.27	0.86	1.04	1.85	2.88	2.26	3.02

All RMSD values for 1U19 and 3ODU (upper left corner) after applying the fragmental GPCR approach on different template structures (first column, grey shaded). The lowest RMSD values for each helix, model and pocket is marked red. The lowest energy model is highlighted in yellow.

Our final models are improved in about half of all test cases and the decrease of the RMSD values is thereby much stronger on average (C_α-Model: 0.27Å, C_α-Pocket: 0.16Å, Pocket: 0.21Å) than the corresponding increase (0.11Å, 0.18Å, 0.11Å, respectively). The biggest success was achieved in modeling 3EML (C_α-RMSD decreased from 2.81Å to 2.44Å) and 3PBL (Pocket-RMSD decreased from 2.88Å to 2.37Å). The improvement of all models with a C_α-RMSD value greater than 2.0Å when applying our approach indicates that it has a high potential for further applications, bearing in mind that we used a very simple energy function and excluding side

Table 6.3: RMSD values of the fragmental GPCR approach

1U19	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
2RH1	1.50	0.96	0.91	1.14	0.96	1.37	1.70	2.36	2.37	3.05
2VT4	1.37	0.99	0.93	1.34	1.03	1.13	1.39	2.17	2.12	2.72
3EML	1.33	1.69	1.53	1.57	1.53	1.71	1.11	2.64	2.69	3.33
3ODU	1.34	2.53	1.05	3.01	1.19	1.05	1.59	2.46	2.47	3.28
3PBL	0.82	0.85	0.66	0.75	0.80	0.95	1.00	1.87	2.03	2.77
3RZE	1.36	1.27	0.91	2.13	0.87	0.96	1.35	2.08	1.65	2.37

3ODU	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.45	2.14	1.40	3.15	1.28	1.10	1.58	2.98	2.60	3.24
2RH1	1.60	2.30	0.61	2.80	0.94	1.30	1.52	2.76	2.92	3.43
2VT4	1.63	2.12	0.75	2.99	1.61	1.77	1.41	3.13	3.11	3.85
3EML	1.54	2.32	1.67	3.01	2.44	1.66	1.69	3.64	3.63	4.28
3PBL	1.45	1.99	0.79	3.01	0.69	1.48	1.38	2.64	2.40	2.89
3RZE	2.28	2.20	0.71	3.19	1.10	1.30	1.96	3.10	2.41	3.10

All RMSD values for 1U19 and 3ODU (upper left corner) after applying the fragmental GPCR approach on different template structures (first column, grey shaded). The lowest RMSD values for each helix, model and pocket is marked red. The lowest energy model is highlighted in yellow.

Table 6.4: RMSD values of the final models obtained by both approaches

Rigid	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	0.84	0.87	0.79	0.77	0.84	0.94	1.08	2.18	2.16	2.90
2RH1	0.90	0.64	0.46	0.71	0.63	0.55	0.79	0.95	0.81	1.69
2VT4	0.45	0.64	0.59	0.73	0.61	0.65	0.65	0.83	0.73	1.48
3EML	1.98	2.02	1.64	2.85	1.73	2.01	1.32	2.81	2.54	3.21
3ODU	1.59	1.96	0.79	3.15	0.69	1.42	1.55	2.78	2.31	2.81
3PBL	1.09	0.71	0.62	1.61	0.76	1.09	1.13	1.51	1.75	2.88
3RZE	1.23	1.01	0.91	1.74	0.68	0.94	0.85	1.65	1.70	2.57

Hinges	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	0.82	0.85	0.66	0.75	0.80	0.95	1.00	1.87	2.03	2.77
2RH1	0.93	0.63	0.46	0.68	0.62	0.52	0.85	1.05	0.94	1.58
2VT4	0.45	0.66	0.59	0.72	0.71	0.66	0.67	0.87	0.89	1.56
3EML	1.96	1.17	1.91	2.06	1.74	2.26	1.07	2.44	2.28	3.14
3ODU	1.45	1.99	0.79	3.01	0.69	1.48	1.38	2.64	2.40	2.89
3PBL	0.86	0.81	0.97	1.50	0.84	1.05	1.05	1.69	1.65	2.37
3RZE	1.57	1.05	0.76	1.75	0.68	0.95	0.91	1.77	1.99	2.75

The RMSD values of the rigid helix approach (top) and our new fragmental modeling approach (bottom) in comparison. The improved values are colored red. Where the values did not change they are colored dark red.

chain optimization. The promising result is similar with regard to individual helices. Although the C_α-RMSD values of some helices became slightly worse, we are able to model a few helices significantly better. For example, H2 and H4 of 3EML decreased from 2.02Å to 1.17Å and from 2.85Å to 2.06Å, respectively. The obvious RMSD value decrease of these helices and the only slight increase of already well modeled helices shows again that our approach is working very well and that the energy function remains the only bottleneck.
The success of our new approach is further highlighted when comparing the C_α-RMSD values of individual helices to those of other approaches. Vaidehi and coworkers stated that they achieved the following results for their bovine rhodopsin model using MembStruk: 1.0Å for H1, 2.1Å for H2, 1.2Å for H3, 1.1Å for H4, 1.8Å for H5, 2.2Å for H6, and 1.6Å for H7.[169] Hence, more than half of their helices have a C_α-RMSD value larger than 1.5Å compared to the corresponding crystal structure. This can be easily improved by far when using artificial evolution to generate initial models and even more when applying our fragmental modeling approach for their optimization. Overall, we were able to model more than 60% of all helices with an C_α-RMSD value smaller or equal to 1.0Å as summarized in Table 6.5. In particular for bovine rhodopsin all modeled helices have a C_α-RMSD value of at most 1.0Å, which corresponds to the lowest C_α-RMSD achieved by Vaidehi et al.

Table 6.5: C_α-RMSD values for individual helices

	$x \leqslant 1.0$	$1.0 < x \leqslant 1.5$	$x > 1.5$
Rigid	29	8	12
Hinges	31	9	9

The C_α-RMSD values for individual helices. About 63% of all helices are modeled with an C_α-RMSD value smaller or equal to 1.0Å. A small improvement was achieved when using hinges.

6.4 CONCLUSION

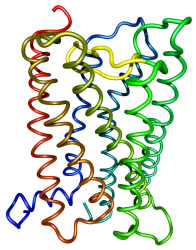
In this chapter, we demonstrated how G-protein coupled receptors can be automatically modeled with a high success rate. Our fragmental modeling approach consists of four steps. First, identify the most conserved residue according to Ballesteros and Weinstein in the template and target sequence. Second, apply the simplified artificial evolution algorithm and SCWRL to obtain an initial model of the target. Third, annotate (randomly) hinges in the outer membrane side of the model. Fourth, optimize the structure using our simulated annealing algorithm followed by a gradient-based optimization procedure. And finally, select the model with the lowest energy value.

The approach we developed is not an ab initio approach in its basic sense since we deduce information from the available template structure. On the other hand, it is also not a pure homology modeling approach because we do not rely on any

alignment, which is said to be the most important step in homology modeling. In our approach, it is sufficient to identify the most conserved residue of each helix and use this as an anchor point. The high structural similarity of the overall fold and of individual helices in all available crystal structures, which belong to different subfamilies, allows us to assume that there will be no surprising difference in one of the unknown rhodopsin-like GPCRs with regard to the transmembrane region. With our new method, we were able to model rhodopsin in this region more precisely (C_α-RMSD: 1.87Å) than any published model from another approach like TASSER (2.1Å),[29] MembStruk (2.8Å)[31] or PREDICT (3.87Å).[32] More important, even the binding pocket residues are modeled in six of the seven cases with an RMSD value smaller than 3.0Å, which can easily be improved when using a more sophisticated energy function.

With our fragmental modeling approach we found a tradeoff between computationally intensive processes like molecular dynamics simulations and too simple optimization procedures where each helix is treated as a single rigid body. Moreover, it is easy to combine different helices from different templates, e. g., based on sequence similarity. Already in 2009, Krause et al. showed that the seven helices should be modeled using different templates.[170] In contrast to other implementations using multiple template homology modeling as MODELLER, our new approach allows easily to choose a template for each helix separately.

7 Conclusion



In this thesis, we analyzed to which extent automated in silico protein modeling strategies can be successfully applied to G-protein coupled receptors without manual interactions by the user. To this end, we designed and implemented a framework for automated GPCR modeling, integrating our own novel techniques as well as state-of-the-art results from the literature.

In our first project, the multiple template GPCR homology modeling approach (see Chapter 3), we used the well-established tool MODELLER to generate a model of the human neurokinin-1 (NK1) receptor based on two templates, bovine rhodopsin (PDB ID: 1F88) and the human β_2 adrenergic receptor (PDB ID: 2RH1). Our model was successfully validated using the technique of virtual screening. Although it was not necessary to modify the templates, e. g., by rotation of a single helix, or to restrain important side chains to specific positions in the modeling procedure, we had to refine at least the sequence alignment and had to select the final model based on expert knowledge and experimental data. Hence, a fully automated modeling procedure was not yet feasible. Our results, however, already showed that additional backbone flexibility can improve the models and that the transmembrane region of GPCRs seems to be extremely similar throughout the whole family.

Our second project was the reimplementation of the promising ab initio algorithm PREDICT developed by Shacham and coworkers (see Chapter 4). Here, we exposed many bottlenecks and difficulties in almost all modeling steps, in particular, in the generation of appropriate initial models. To apply this approach to model unknown GPCRs will probably always fail, if the user does not strongly interact in the modeling procedure. Allowance must be made for the fact that PREDICT, as well as other ab initio modeling approaches, have been developed based on the knowledge of only a single GPCR crystal structure, namely bovine rhodopsin. Much effort has therefore been put into predicting a suitable initial conformation based on the protein's sequence. However, a comparison of all currently available native structures shows a high similarity of the transmembrane region, and hence all crystal structure can be serve as an initial conformation. We also tried to predict the orientation and tilt angle of helices using various methods. However, despite all improvements, no satisfying procedure could be established.

The focus in modeling GPCRs has completely changed with the publication of the native structure of the human CXCR4 chemokine receptor in 2010. This receptor is evolutionarily not related to those with previously known structure but the TM region is still very similar. A recent review of Katritch and co-workers quantified the structural similarity of the seven available GPCRs.[171] They concluded that the intracellular domain is strongly conserved throughout the GPCR family but undergoes

larger changes upon receptor activation. In contrast, the outer membrane domain including the binding pocket is more diverse but changes only slightly during ligand binding.

These observations provide the basis of our new fragmental GPCR modeling algorithm as described in Chapter 6. The outer membrane side is treated very flexibly but with a sufficiently small number of degrees of freedom to allow the algorithm to be still very fast. The more conserved intracellular region, on the other hand, is represented as one rigid cylinder that is only slightly rotated and translated during the optimization process. Applying this method on all available crystal structures, we were able to generate top ranked models with low RMSD values within the transmembrane region. However, the main improvement compared to other approaches is due to the information we deduced from the newly published crystal structures. Instead of predicting suitable initial models, we used a simplified artificial evolution algorithm to convert each helix of the template to one of the target. Because our initial models have been in an appropriate conformation, we neglected side chain optimization and used the side chain positions computed by SCWRL. Hence, our algorithm was not only applied successfully but is also very fast. Since we do not rely on any sequence alignment, this approach cannot be regarded as homology modeling. Instead, it is a kind of hybrid approach that can use any generated helix, e. g., from a molecular dynamics simulation, and map it onto an arbitrary known GPCR fold.

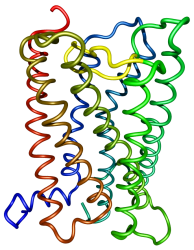
As a side-effect of our work on automated GPCR modeling, we also developed a highly accurate technique for predicting distortions from optimal α-helical geometry based on the corresponding sequence. We, as well as other researchers, are convinced that information about kinks is not completely locally encoded in the sequence. As inter-helical interactions cannot be neglected, it is most probable that our prediction accuracy cannot be significantly increased in future projects if only taking the sequence in the near neighborhood into account. Whereas in former studies helices have been treated as rigid cylinders, our approach is the first algorithm that is able to use this information and allow helices form kinks during the optimization efficiently.

Concluding our research, we can confirm that it is possible to automatically model GPCRs with regard to their binding pockets with a surprisingly high precision. Based on the first modeling algorithm, which accounts for the structural diversity of GPCRs, i. e., modeling the outer membrane side very flexibly, while treating the intracellular domain more rigidly, we made remarkable progress towards automated GPCR modeling. One key element is to take advantage of the high structural similarity of the most conserved residues, which serves as an anchor in our modeling procedure. Compared to the recent study of Sander et al., who focused on a very general approach of protein modeling and obtained GPCR models of at least 4Å, approaches particularly tailored to the GPCR case obtain better results. However, the work of Sander can help to improve our kink prediction method as it is able to identify the amino acids that have spatial contacts. Thus, to include this information might improve our kink prediction results.

Moreover, we were able to always identify the best template structure for each target. Since template selection is not always straightforward, as it was demonstrated in a recent study by Zhu and Li on the β1-adrenoceptor receptor, our approach can be used as a preprocessing step for homology modeling to find the most suitable template structure.

Although protein modeling in general is not solved yet, and it might take still some decades, we made large steps towards automated GPCR modeling. The idea of treating helices as a collection of rigid fragments to cope with kinks is new in protein modeling and is for sure also applicable for other proteins consisting mainly of helices. Since our approach can also be used for template selection in homology modeling, it is quite obvious that it is very useful for various applications and thus we hope that it will be of use in many exciting projects in the future. Furthermore, the modular design enables the user to use our program for other optimization problems of rigid fragments, we don't have in mind yet.

A *Appendix*



The appendix contains several tables and figures from the results of the methods we applied on all seven available crystal structures.

Tables:

Figures:

Table A.1: SSP results of TMPRED

1U19	H1	H2	H3	H4	H5	H6	H7
first residue	I.32	II.41	III.30	IV.42	V.38	VI.36	VII.27
last residue	I.58	II.66	III.55	IV.64	V.59	VI.57	VII.56
helix length	27	26	26	23	22	22	30

2RH1	H1	H2	H3	H4	H5	H6	H7
first residue	I.37	II.41	III.27	IV.44	V.37	VI.36	VII.33
last residue	I.57	II.67	III.48	IV.64	V.59	VI.57	VII.56
helix length	21	27	22	21	23	22	24

2VT4	H1	H2	H3	H4	H5	H6	H7
first residue	I.33	II.41	III.27	IV.42	V.37	VI.36	VII.33
last residue	I.57	II.66	III.48	IV.65	V.59	VI.57	VII.56
helix length	25	26	22	24	23	22	24

3EML	H1	H2	H3	H4	H5	H6	H7
first residue	I.34	II.41	III.20	IV.44	V.38	VI.36	VII.32
last residue	I.59	II.68	III.48	IV.64	V.59	VI.57	VII.55
helix length	26	28	29	21	22	22	24

3ODU	H1	H2	H3	H4	H5	H6	H7
first residue	I.33	II.44	III.27	IV.44	V.43	VI.36	VII.34
last residue	I.57	II.71	III.48	IV.64	V.63	VI.57	VII.56
helix length	25	28	22	21	21	22	23

3PBL	H1	H2	H3	H4	H5	H6	H7
first residue	I.33	II.41	III.20	IV.42	V.38	VI.36	VII.33
last residue	I.57	II.66	III.48	IV.64	V.65	VI.57	VII.56
helix length	25	26	29	23	28	22	24

3RZE	H1	H2	H3	H4	H5	H6	H7
first residue	I.32	II.41	III.27	IV.42	V.37	VI.36	VII.32
last residue	I.57	II.66	III.48	IV.63	V.59	VI.57	VII.52
helix length	26	26	22	22	23	22	21

The results of TMPRED given in Ballesteros-Weinstein nomenclature. Predicted helical ends where binding pocket residues are missed are marked red.

Table A.2: SSP results of TMHMM

1U19	H1	H2	H3	H4	H5	H6	H7
first residue	I.34	II.41	III.26	IV.42	V.37	VI.37	VII.33
last residue	I.56	II.63	III.48	IV.64	V.59	VI.59	VII.55
helix length	23	23	23	23	23	23	23

2RH1	H1	H2	H3	H4	H5	H6	H7
first residue	I.35	II.41	III.26	IV.42	V.40	VI.37	VII.34
last residue	I.57	II.63	III.48	IV.61	V.62	VI.59	VII.53
helix length	23	23	23	20	23	23	20

2VT4	H1	H2	H3	H4	H5	H6	H7
first residue	I.36	II.51	III.36	IV.44	V.37	VI.36	VII.33
last residue	I.58	II.73	III.58	IV.66	V.59	VI.58	VII.55
helix length	23	23	23	23	23	23	23

3EML	H1	H2	H3	H4	H5	H6	H7
first residue	I.35	II.42	III.24	IV.42	V.41	VI.38	VII.33
last residue	I.57	II.64	III.46	IV.64	V.63	VI.60	VII.55
helix length	23	23	23	23	23	23	23

3ODU	H1	H2	H3	H4	H5	H6	H7
first residue	I.37	II.44	III.27	IV.44	V.40	VI.36	VII.34
last residue	I.59	II.62	III.48	IV.63	V.62	VI.55	VII.56
helix length	23	19	22	20	23	20	23

3PBL	H1	H2	H3	H4	H5	H6	H7
first residue	I.35	II.42	III.26	IV.42	V.42	VI.37	VII.38
last residue	I.57	II.64	III.48	IV.64	V.64	VI.59	VII.60
helix length	23	23	23	23	23	23	23

3RZE	H1	H2	H3	H4	H5	H6	H7
first residue	I.34	II.41	III.26	IV.42	V.40	VI.37	VII.30
last residue	I.56	II.63	III.48	IV.64	V.62	VI.59	VII.52
helix length	23	23	23	23	23	23	23

The results of TMHMM given in Ballesteros-Weinstein nomenclature. Predicted helical ends where binding pocket residues are missed are marked red.

Table A.3: SSP results of PHDhtm

1U19	H1	H2	H3	H4	H5	H6	H7
first residue	I.30	II.41	III.25	IV.42	V.38	VI.36	(VII.34)
last residue	I.58	II.67	III.53	IV.64	V.64	(VI.55)	VII.56
helix length	29	27	29	23	27	20	23

2RH1	H1	H2	H3	H4	H5	H6	H7
first residue	I.33	(II.43)	III.23	IV.42	V.37	VI.37	VII.33
last residue	(I.56)	II.68	III.53	IV.62	V.63	VI.61	VII.53
helix length	24	26	31	21	27	25	21

2VT4	H1	H2	H3	H4	H5	H6	H7
first residue	I.36	(II.42)	III.23	IV.43	V.35	VI.36	VII.33
last residue	(I.57)	II.66	III.54	IV.62	V.62	VI.61	VII.57
helix length	22	25	31	20	28	26	25

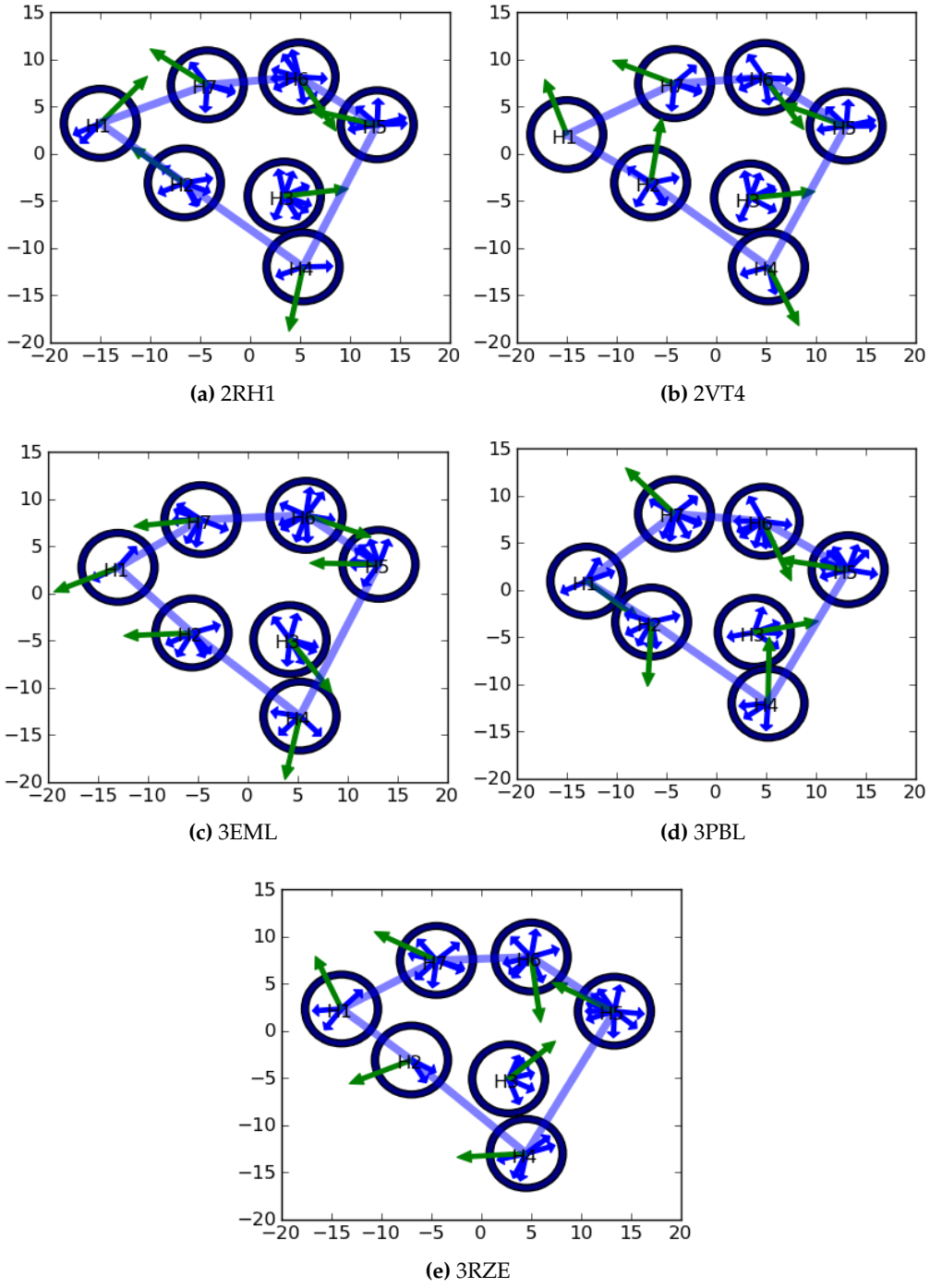
3EML	H1	H2	H3	H4	H5	H6	H7
first residue	I.30	II.37	(III.20)	IV.43	V.32	VI.36	VII.31
last residue	I.58	(II.65)	III.54	IV.61	V.65	VI.59	VII.54
helix length	29	29	35	19	34	24	24

3ODU	H1	H2	H3	H4	H5	H6	H7
first residue	I.32	II.43	(III.28)	IV.44	V.37	VI.34	VII.38
last residue	I.59	(II.64)	III.54	IV.62	V.64	VI.59	VII.55
helix length	28	22	27	19	28	25	18

3PBL	H1	H2	H3	H4	H5	H6	H7
first residue	I.32	II.40	III.24	IV.43	V.37	VI.35	VII.33
last residue	I.58	II.66	III.53	IV.62	V.63	VI.60	VII.56
helix length	27	27	30	20	27	26	24

3RZE	H1	H2	H3	H4	H5	H6	H7
first residue	I.33	II.38	(III.27)	IV.42	V.36	VI.36	VII.27
last residue	I.59	(II.64)	III.53	IV.63	V.62	VI.63	VII.51
helix length	27	27	27	22	27	28	25

The results of PHDThtm given in Ballesteros-Weinstein nomenclature. In some cases, PHDhtm predicted one very long instead of two single helices. Here, we used the refinement method (PHDThtm) and set them in brackets. Predicted helical ends where binding pocket residues are missed are marked red.

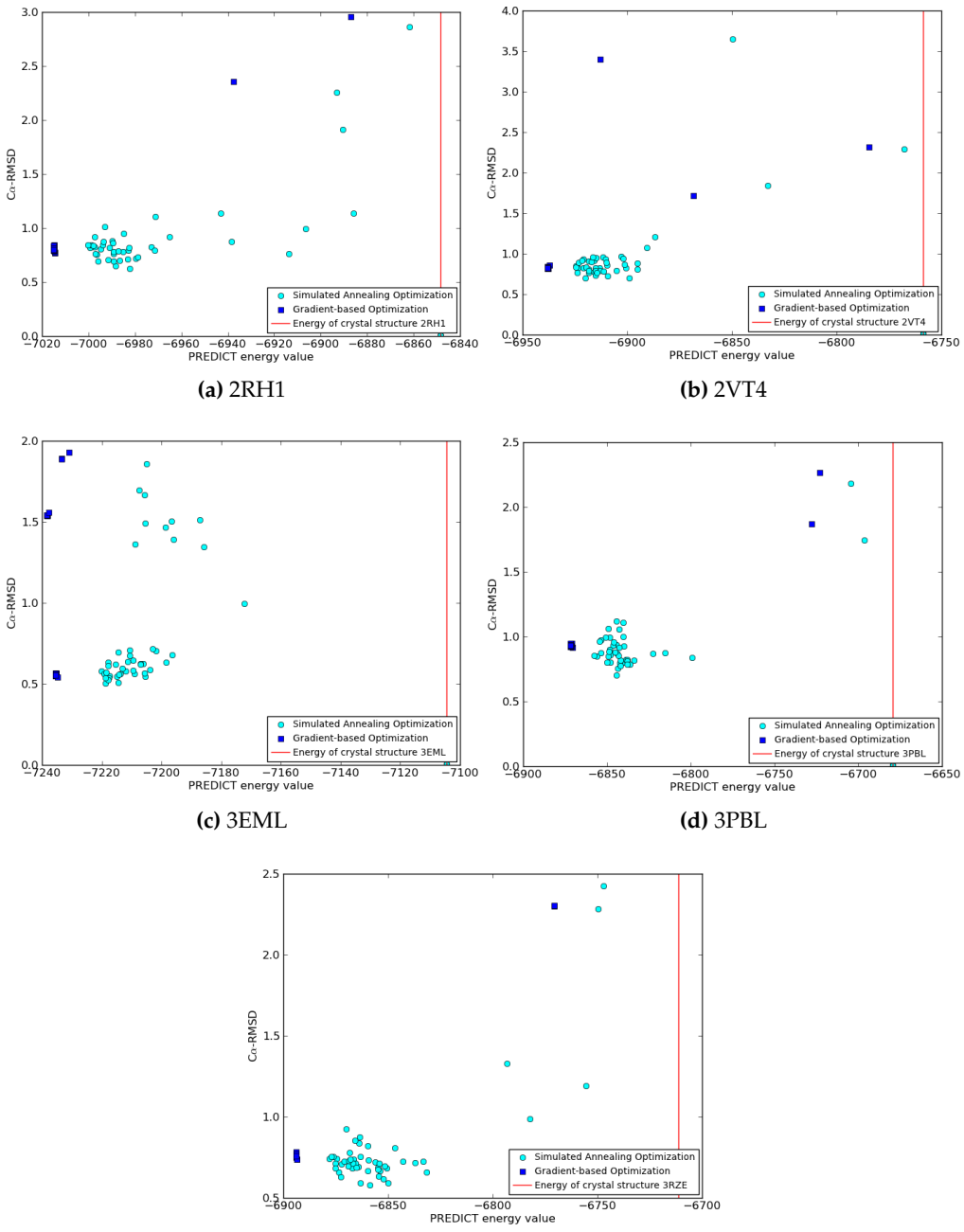


(a) 2RH1 **(b)** 2VT4

(c) 3EML **(d)** 3PBL

(e) 3RZE

Figure A.1: The canonical version of native structures in 2D. The helices in the crystal structure were replaced by their canonical version. The x,y-center is the average of the backbone atoms. The hydrophobic moment (green arrow) is computed using the hydrophobicity scale of Eisenberg and trapezoidal weighting factors. All inter-helical vectors are drawn as blue arrows.



(a) 2RH1 **(b)** 2VT4

(c) 3EML **(d)** 3PBL

(e) 3RZE

Figure A.2: 50 unrestrained SA runs each followed by gradient- based optimization applied to the crystal structure conformation.

Table A.4: RMSD values of the rigid GPCR approach

2RH1	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.51	0.81	0.85	1.14	1.15	1.07	1.56	2.08	2.00	2.76
2VT4	0.90	0.64	0.46	0.71	0.63	0.55	0.79	0.95	0.81	1.69
3EML	1.58	1.55	1.45	1.45	1.48	1.25	1.26	2.31	2.13	2.80
3ODU	2.04	2.53	0.73	2.68	0.83	1.07	1.50	2.98	2.78	3.69
3PBL	1.32	0.85	0.65	1.29	0.61	1.03	1.34	1.85	1.54	2.37
3RZE	1.55	1.00	0.78	2.18	0.80	1.13	1.38	1.98	1.42	2.40

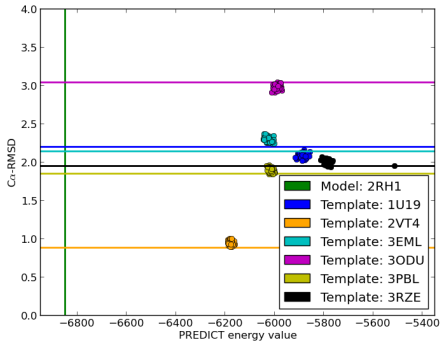
2VT4	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.55	0.92	0.90	1.19	1.14	1.06	1.52	2.13	2.19	2.79
2RH1	0.45	0.64	0.59	0.73	0.61	0.65	0.65	0.83	0.73	1.48
3EML	1.00	1.55	1.47	1.78	1.62	1.38	1.22	2.47	2.47	3.18
3ODU	1.92	2.37	0.73	2.92	0.87	1.25	1.40	2.91	2.84	3.10
3PBL	1.44	0.74	0.61	1.45	0.58	1.01	1.19	1.80	1.47	2.11
3RZE	1.51	0.96	0.83	2.48	0.82	1.11	1.41	1.98	1.66	2.86

3EML	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.67	1.57	1.42	1.52	1.72	2.09	1.00	2.42	2.44	3.14
2RH1	1.51	1.29	1.58	1.59	1.70	1.66	1.33	2.15	2.15	2.75
2VT4	1.60	1.26	1.44	1.66	1.67	1.54	1.19	2.28	2.30	3.42
3ODU	1.98	2.02	1.64	2.85	1.73	2.01	1.32	2.81	2.54	3.21
3PBL	1.41	1.33	1.28	1.61	1.64	2.34	0.65	1.98	1.77	2.55
3RZE	1.83	1.33	1.74	1.88	1.78	1.98	0.90	2.26	1.92	2.86

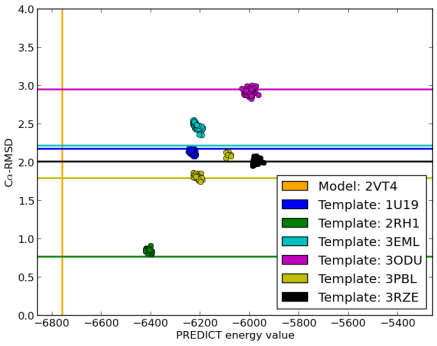
3PBL	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.06	0.79	0.84	0.65	0.80	0.98	1.13	1.70	1.79	2.50
2RH1	0.97	0.85	0.57	1.19	0.85	1.14	1.35	1.49	1.56	2.37
2VT4	1.09	0.71	0.62	1.61	0.76	1.09	1.13	1.51	1.75	2.88
3EML	0.72	1.46	1.18	1.52	1.64	1.79	0.50	1.82	1.91	2.76
3ODU	1.54	2.22	0.81	2.71	0.80	1.28	1.49	2.27	2.18	3.28
3RZE	0.97	0.92	0.98	1.67	0.78	0.93	1.04	1.42	1.34	2.35

3RZE	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.54	1.32	1.23	1.74	0.85	0.78	1.58	2.00	2.08	2.79
2RH1	1.57	0.99	0.68	1.98	0.83	1.04	1.31	1.87	1.80	2.69
2VT4	1.54	1.07	0.68	2.46	0.86	1.07	1.23	1.99	1.98	2.82
3EML	1.11	1.39	1.72	1.91	1.89	1.57	1.18	2.26	2.36	3.63
3ODU	1.90	2.13	0.76	2.89	0.86	1.09	1.68	2.26	2.16	3.60
3PBL	1.23	1.01	0.91	1.74	0.68	0.94	0.85	1.65	1.70	2.57

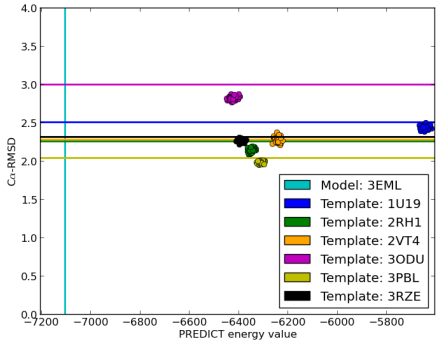
All RMSD values for different structures (upper left corner) obtained by the rigid helix approach. The lowest RMSD values for each helix, model and pocket is marked red. The lowest energy model is highlighted in yellow.



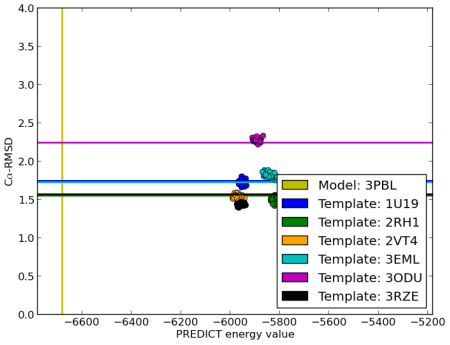
(a) 2RH1



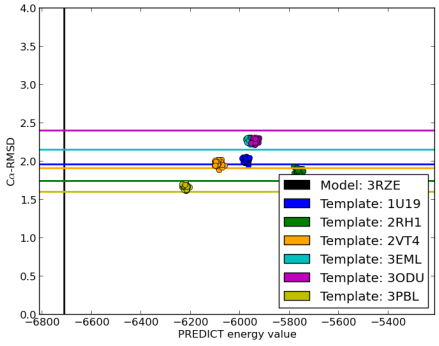
(b) 2VT4



(c) 3EML



(d) 3PBL



(e) 3RZE

Figure A.3: 50 restrained single rigid body SA runs using the side chain positions computed by SCWRL4 applied to the initial conformation created by AE. The horizontal lines represent the C_α-RMSD values of the corresponding initial conformation. The vertical line is the energy value of the crystal structure.

Table A.5: RMSD values of the fragmental GPCR approach

2RH1	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.46	0.77	0.92	1.10	1.39	1.31	1.55	2.30	2.20	2.81
2VT4	0.93	0.63	0.46	0.68	0.62	0.52	0.85	1.05	0.94	1.58
3EML	1.47	1.43	1.41	1.41	1.54	1.12	1.24	2.67	2.62	3.12
3ODU	1.34	2.53	1.05	3.01	1.19	1.05	1.59	2.46	2.47	3.28
3PBL	1.27	0.78	0.65	1.33	0.61	1.73	1.28	2.51	2.07	2.90
3RZE	1.36	0.85	0.84	2.23	0.80	1.27	1.27	1.89	1.45	2.33

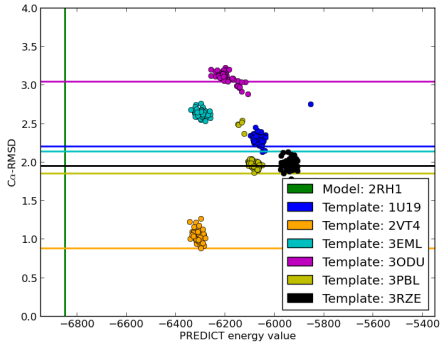
2VT4	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.28	0.85	0.86	1.24	1.26	1.13	1.50	2.17	2.26	2.82
2RH1	0.45	0.66	0.59	0.72	0.71	0.66	0.67	0.87	0.89	1.56
3EML	1.21	1.46	1.43	1.61	1.82	1.39	1.19	2.67	2.81	3.48
3ODU	2.02	2.16	0.71	2.92	0.89	1.65	1.88	3.07	2.97	3.17
3PBL	1.45	0.75	1.72	1.54	0.66	1.19	1.12	2.48	2.36	2.82
3RZE	1.60	0.76	0.86	2.53	0.82	1.37	1.43	2.32	2.05	3.11

3EML	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.85	1.43	1.77	1.45	1.84	2.72	0.90	2.74	2.65	3.42
2RH1	1.49	1.28	1.51	1.71	1.69	2.93	1.28	3.05	3.02	3.58
2VT4	1.56	0.90	1.27	1.67	1.67	1.76	1.23	1.99	1.87	2.98
3ODU	1.91	1.94	1.95	2.82	1.63	2.18	1.33	2.93	2.42	3.15
3PBL	1.49	1.48	1.32	1.40	1.67	2.92	0.65	2.79	2.39	2.99
3RZE	1.96	1.17	1.91	2.06	1.74	2.26	1.07	2.44	2.28	3.14

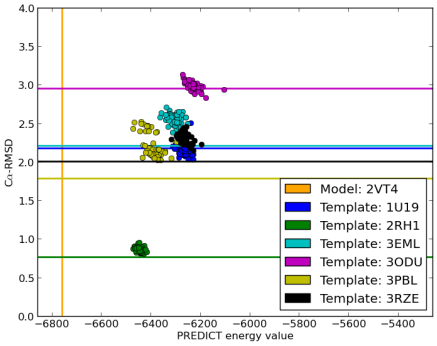
3PBL	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.01	0.77	0.84	0.65	1.01	0.86	1.54	1.85	1.94	2.46
2RH1	0.90	0.98	0.57	1.36	0.98	1.19	1.37	1.53	1.60	2.40
2VT4	0.86	1.02	0.65	1.71	0.79	1.10	1.16	1.62	1.77	2.78
3EML	0.73	1.29	1.26	1.31	1.79	1.52	0.52	1.91	2.00	2.83
3ODU	1.54	1.92	0.99	2.75	0.84	1.41	2.22	2.68	2.82	3.86
3RZE	0.86	0.81	0.97	1.50	0.84	1.05	1.05	1.69	1.65	2.37

3RZE	H1	H2	H3	H4	H5	H6	H7	C_α-Model	C_α-Pocket	Pocket
1U19	1.56	1.08	1.44	1.54	1.01	0.76	1.78	2.05	2.22	2.88
2RH1	1.44	1.01	0.72	2.06	0.97	1.01	1.63	2.08	2.18	2.93
2VT4	1.42	1.06	0.80	2.37	0.86	1.02	1.40	2.02	2.02	2.83
3EML	1.21	1.14	2.14	2.19	1.96	1.58	1.22	2.56	2.76	3.89
3ODU	2.13	2.02	0.97	3.03	0.73	1.11	2.13	2.77	2.88	4.05
3PBL	1.57	1.05	0.76	1.75	0.68	0.95	0.91	1.77	1.99	2.75

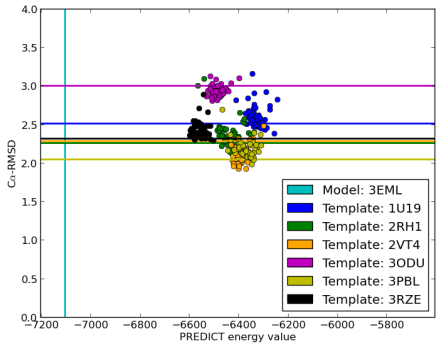
All RMSD values for different structures (upper left corner) obtained by the fragmental GPCR modeling approach. The lowest RMSD values for each helix, model and pocket is marked red. The lowest energy model is highlighted in yellow.



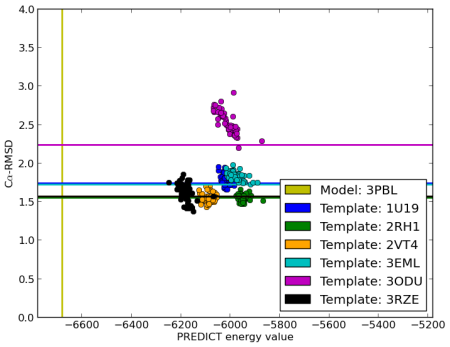
(a) 2RH1



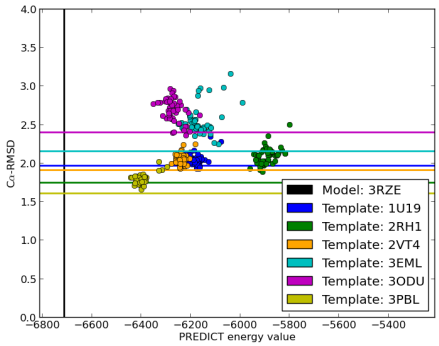
(b) 2VT4



(c) 3EML

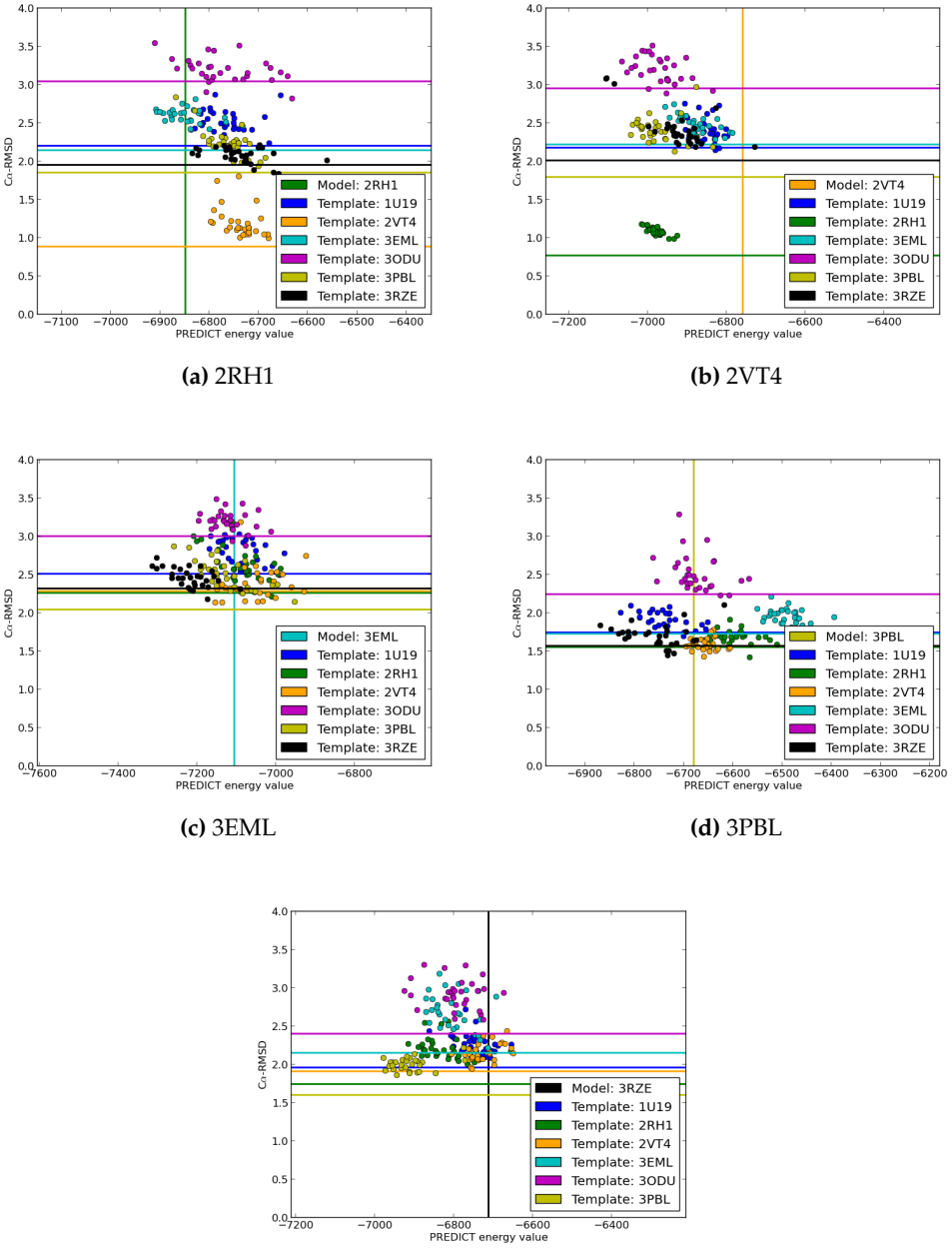


(d) 3PBL



(e) 3RZE

Figure A.4: 50 restrained fragmental modeling SA runs using the side chain positions computed by SCWRL4 applied to the initial conformation created by AE. The horizontal lines represent the C_α-RMSD values of the corresponding initial conformation. The vertical line is the energy value of the crystal structure.

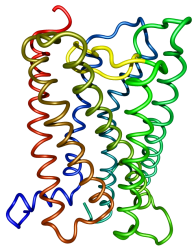


(a) 2RH1 **(b)** 2VT4

(c) 3EML **(d)** 3PBL

(e) 3RZE

Figure A.5: 50 unrestrained SA runs including side chain optimization applied to the initial conformation created by AE. The horizontal lines represent the C_α-RMSD values of the corresponding initial conformation. The vertical line is the energy value of the crystal structure.



B *Copyrights*

Some figures in this thesis are protected by copyright, and thus license numbers for previously published content have been acquired through the Copyright Clearance Center. In the following list, we give the original sources and the license number where appropriate.

- Figure 1.1 by Yekaterina Kadyshevskaya is used with kind permisson by Angela L. Walker, GPCR Network Program Manager.
- Figure 2.1 is used with permission from *Journal of Cell Science: jcs.biologists.org*: J.-P. Vilardaga, L. F. Agnati, K. Fuxe, and F. Ciruela. G-protein-coupled receptor heteromer dynamics., 24(Pt 24):4215–4220, **2010**.
- Figure 2.2 is used with permission from *Nature Reviews Cancer*: R. T. Dorsam and J. S. Gutkind. G-protein-coupled receptors and cancer., 7:79–94, **2007**. License number: 2958790379588
- Figure 2.3 and 2.4 are used with permission from *American Society for Biochemistry and Molecular Biology*.
- Figure 4.2 is used with permission from *Proteins: Structure, Function and Bioinformatics*: J. S. Surgand, J. Rodrigo, E. Kellenberger, and D. Rognan. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors., 62(2):509–538, **2006**. License number: 2958810380166
- Figure 4.5b is used with kind permisson by Dr. Alpeshkumar Malde, School of Chemistry and Molecular Biology, University of Queensland.
- Figure 4.6 and 4.9 are used with permission from *Proteins: Structure, Function and Bioinformatics*: S. Shacham, Y. Marantz, S. Bar-Haim, O. Kalid, D. Warshaviak, N. Avisar, B. Inbal, A. Heifetz, M. Fichman, M. Topf, Z. Naor, S. Noiman, and O. M. Becker. PREDICT modeling and in silico screening for G-protein coupled receptors., 57(1):51–86, **2004**. License number: 2958811393692
- Figure 4.15 has no licensing restrictions.
- All figures in chapter 3 and 5 were previously published in our own manuscripts,[77,144] and the permission is granted by the *Journal of Medicinal Chemistry* and the *Journal of Chemical Information and Modeling* in both print and electronic formats.

BIBLIOGRAPHY

[1] J. C. Kendrew, G. Bodo, H. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, **1985**.

[2] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112(3):535–542, **1977**.

[3] S. Jayasinghe, K. Hristova, and S. H. White. MPtopo: A database of membrane protein topology. *Protein Sci.*, 10(2):455–458, **2001**.

[4] M. Williamson, T. Havel, and K. Wüthrich. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.*, 182(2):295–315, **1985**.

[5] C. Huang and S. Mohanty. Challenging the limit: NMR Assignment of a 31 kDa Helical Membrane Protein. *J. Am. Chem. Soc.*, 132(11):3662–3663, **2010**.

[6] D. M. Rosenbaum, S. G. F. Rasmussen, and B. K. Kobilka. The structure and function of G-protein-coupled receptors. *Nature*, 459(7245):356–363, **2009**.

[7] M. J. Serrano-Vega, F. Magnani, Y. Shibata, and C. G. Tate. Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. *Proc. Natl. Acad. Sci.*, 105(3):877–882, **2009**.

[8] D. M. Rosenbaum, V. Cherezov, M. A. Hanson, S. G. F. Rasmussen, F. S. Thian, T. S. Kobilka, H.-J. Choi, X.-J. Yao, W. I. Weis, R. C. Stevens, and B. K. Kobilka. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science*, 318(5854):1266–1273, **2007**.

[9] V. Cherezov, D. M. Rosenbaum, M. A. Hanson, S. G. Rasmussen, F. S. Thian, T. S. Kobilka, H. J. Choi, P. Kuhn, W. I. Weis, B. K. Kobilka, and S. R. C. High Resolution Crystal Structure of an Engineered Human $\mathtt{beta}_2$ Adrenergic G protein-Coupled Receptor. *Science*, 318(5854):1258–1265, **2007**.

[10] T. Warne, M. J. Serrano-Vega, J. G. Baker, R. Moukhametzianov, P. C. Edwards, R. Henderson, A. G. Leslie, C. G. Tate, and G. F. Schertler. Structure of a $\mathtt{beta}_1$-adrenergic G protein-coupled receptor. *Nature*, 454(7203):486–491, **2008**.

[11] V. P. Jaakola, M. T. Griffith, M. A. Hanson, V. Cherezov, E. Y. Chien, J. R. Lane, A. P. Ijzerman, and R. C. Stevens. The 2.6 Angstrom Crystal Structure of a Human A_{2A} Adenosine Receptor Bound to an Antagonist. *Science*, 322(6027):1211–1217, **2008**.

[12] A. Manglik, A. C. Kruse, T. S. Kobilka, F. S. Thian, J. M. Mathiesen, R. K. Sunahara, L. Pardo, W. I. Weis, B. K. Kobilka, and S. Granier. Crystal structure of the mu-opioid receptor bound to a morphinan antagonist. *Nature*, 485(7398):321–326, **2012**.

[13] H. Wu, D. Wacker, M. Mileni, V. Katritch, G. W. W. Han, E. Vardy, W. Liu, A. A. Thompson, X.-P. P. Huang, F. I. Carroll, S. W. Mascarella, R. B. Westkaemper, P. D. Mosier, B. L. Roth, V. Cherezov, and R. C. Stevens. Structure of the human kappa-opioid receptor in complex with JDTic. *Nature*, 485(7398):327–332, **2012**.

[14] A. A. Thompson, W. Liu, E. Chun, V. Katritch, H. Wu, E. Vardy, X.-P. P. Huang, C. Trapella, R. Guerrini, G. Calo, B. L. Roth, V. Cherezov, and R. C. Stevens. Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature*, 485(7398):395–399, **2012**.

[15] S. Granier, A. Manglik, A. C. Kruse, T. S. Kobilka, F. S. Thian, W. I. Weis, and B. K. Kobilka. Structure of the delta-opioid receptor bound to naltrindole. *Nature*, 485(7398):400–404, **2012**.

[16] M. A. Hanson, C. B. Roth, E. Jo, M. T. Griffith, F. L. Scott, G. Reinhart, H. Desale, B. Clemons, S. M. Cahalan, S. C. Schuerer, M. G. Sanna, G. W. W. Han, P. Kuhn, H. Rosen, and R. C. Stevens. Crystal structure of a lipid G protein-coupled receptor. *Science*, 335(6070):851–855, **2012**.

[17] H. B. Schiöth and R. Fredriksson. The GRAFS classification system of G-protein coupled receptors in comparative perspective. *Gen. Comp. Endocrinol*, 142(1-2):94–101, **2005**.

[18] M. J. Gabanyi, P. D. Adams, K. Arnold, L. Bordoli, L. G. Carter, J. Flippen-Andersen, L. Gifford, J. Haas, A. Kouranov, W. A. McLaughlin, D. I. Micallef, W. Minor, R. Shah, T. Schwede, Y. P. Tao, J. D. Westbrook, M. Zimmerman, and H. M. Berman. The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J. Struct. Funct. Genomics*, 12(2):45–54, **2011**.

[19] H. L. Brinks and E. A. D. Regulation of GPCR signaling in hypertension. *Biochim. Biophys. Acta.*, 1802(12):1278–1275, **2010**.

[20] M. S. Lombardi, A. Kavelaars, and H. C. J. Role and modulation of G protein-coupled receptor signaling in inflammatory processes. *Crit. Rev. Immunol.*, 22(2):141–163, **2002**.

[21] D. A. Deshpande and R. B. Penn. Targeting G protein-coupled receptor signaling in asthma. *Cell Signal*, 18(12):2105–2120, **2006**.

[22] T. Kienast and A. Heinz. Dopamine and the diseased brain. *CNS Neurol. Disord. Drug Targets*, 5(1):109–131, **2006**.

[23] B. Contreras-Moreira and P. A. Bates. Domain Fishing: a first step in protein comparative modelling. *Bioinformatics*, 18(8):1141–1142, **2002**.

[24] M. Nielsen, C. Lundegaard, O. Lund, and T. N. Petersen. CPHmodels-3.0 – remote homology modeling using structure-guided sequence profiles. *Nucl. Acids Res.*, 38:W576–581, **2010**.

[25] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using MODELLER. *Curr. Prot. Bioinf*, Chapter 2:Unit 2 9, **2007**.

[26] C. Combet, M. Jambon, G. Deléage, and C. Geourjon. Geno3D: Automatic comparative molecular modelling of protein. *Bioinformatics*, 18(1):213–214, **2002**.

[27] M. Zhu and M. Li. Revisiting the homology modeling of G-protein coupled receptors: beta1-adrenoceptor as an example. *Mol. BioSyst.*, 8(6):1686–1693, **2012**.

[28] Y. Zhang and J. Skolnick. Automated structure prediction of weakly homologous proteins on a genomic scale. *PNAS*, 101(20):7594–7599, **2003**.

[29] Y. Zhang, M. E. DeVries, and J. Skolnick. Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput. Biol.*, 2(2):88–99, **2006**.

[30] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf1, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One*, 6(12):e28766, **2011**.

[31] R. J. Trabanino, S. E. Hall, N. Vaidehi, W. B. Floriano, V. W. Kam, and W. A. Goddard. First Principles Predictions of the Structure and Function of G-protein-coupled-receptors: Validation for Bovine Rhodopsin. *Biophys. J.*, 86(4):1904–1921, **2004**.

[32] S. Shacham, M. Topf, N. Avisar, F. Glaser, Y. Marantz, S. Bar-Haim, S. Noiman, Z. Naor, and O. M. Becker. Modeling the 3D structure of GPCRs from sequence. *Med. Res. Rev.*, 21(5):472–483, **2001**.

[33] S. Shacham, Y. Marantz, S. Bar-Haim, O. Kalid, D. Warshaviak, N. Avisar, B. Inbal, A. Heifetz, M. Fichman, M. Topf, Z. Naor, S. Noiman, and O. M. Becker. PREDICT modeling and in silico screening for G-protein coupled receptors. *Proteins: Struct., Func., Bioinf.*, 57(1):51–86, **2004**.

[34] F. M. McRobb, B. Capuano, I. T. Crosby, D. K. Chalmers, and E. Yuriev. Homology modeling and docking evaluation of aminergic G protein-coupled receptors. *J. Chem. Inf. Mod.*, 50(4):626–637, **2010**.

[35] J. Simms, N. E. Hall, P. H. C. Lam, L. J. Miller, A. Christopoulos, R. Abagyan, and P. M. Sexton. Homology modeling of GPCRs. In W. R. Leifert, editor, *Methods Mol. Biol.*, volume 553, pages 97–113. Springer, **2009**.

[36] P. Herzyk and R. E. Hubbard. A Reduced Representation of Proteins for Use in Restraint Satisfaction Calculations. *Proteins*, 17(3):310–324, **1993**.

[37] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, **2001**.

[38] R. A. Dixon, B. K. Kobilka, D. J. Strader, J. L. Benovic, H. G. Dohlman, T. Frielle, M. A. Bolanowski, C. D. Bennett, E. Rands, R. E. Diehl, R. A. Mumford, E. E. Slater, I. S. Sigal, M. G. Caron, R. J. Lefkowitz, and C. D. Strader. Cloning of the gene and cDNA for mammalian beta-adrenergic receptor and homology with rhodopsin. *Nature*, 321(6065):75–79, **1986**.

[39] http://www.imshealth.com/.

[40] K. McConalogue and N. W. Bunnett. G protein-coupled receptors in gastrointestinal physiology. II. Regulation of neuropeptide receptors in enteric neurons. *Am. J. Physiol.*, 274(5 Pt 1):G792–G796, **2007**.

[41] S. J. Hill. Distribution, properties, and functional characteristics of three classes of histamine receptors. *Pharmacol. Rev.*, 42(1):45–83, **1990**.

[42] K. Fuxe, P. Manger, S. Genedani, and L. Agnati. The nigrostriatal DA pathway and Parkinson's disease. *J. Neural. Transm.*, 70(1):71–83, **2006**.

[43] J.-P. Vilardaga, L. F. Agnati, K. Fuxe, and F. Ciruela. G-protein-coupled receptor heteromer dynamics. *J. Cell. Sci.*, 24(Pt 24):4215–4220, **2010**.

[44] R. T. Dorsam and J. S. Gutkind. G-protein-coupled receptors and cancer. *Nat. Rev. Cancer*, 7:79–94, **2007**.

[45] K. Illergaard, D. H. Ardell, and A. Elofsson. Structure is three to ten times more conserved than sequence – a study of structural response in protein cores. *Proteins*, 77(3):499–508, **2009**.

[46] B. J. A. and H. Weinstein. Integrated Methods for the Construction of Three-Dimensional Models and Computational Probing of Structure-Function Relations in G Protein-Coupled Receptors. *Methods Neurosci.*, 25:366–428, **1995**.

[47] F. Z. Chung, C. D. Wang, P. C. Potter, J. C. Venter, and C. M. Fraser. Site-directed mutagenesis and continuous expression of human beta-adrenergic receptors. Identification of a conserved aspartate residue involved in agonist binding and receptor activation. *J. Biol. Chem.*, 263(9):4052–4055, **1988**.

[48] M. M. Rosenkilde, K. T. N., and T. W. Schwartz. High constitutive activity of a virus-encoded seven transmembrane receptor in the absence of the conserved DRY motif (Asp-Arg-Tyr) in transmembrane helix 3. *Mol. Pharmacol.*, 68(1):11–19, **2005**.

[49] J. Gripentrog, A. Jesaitis, and H. M. Miettinen. A single amino acid substitution (N297A) in the conserved NPXXY sequence of the human N-formyl peptide receptor results in inhibition of desensitization and endocytosis, and a dose-dependent shift in p42=44 mitogen-activated protein kinase activation and chemotaxis. *Biochem. J.*, 352 Pt 2:399–407, **2000**.

[50] U. Gether and B. K. Kobilka. G Protein-coupled Receptors: II. Mechanism of agonist activation. *J. Biol. Chem.*, 273:17979–17982, **1998**.

[51] L. Shi and J. A. Javitch. The Binding Site of Aminergic G Protein-Coupled Receptors: The Transmembrane Segments and Second Extracellular Loop. *annu. Rev. Pharmacol. Toxicol.*, 42:437–467, **2002**.

[52] D. Massotte and B. L. Kiefer. The second extracellular loop: a damper for G protein-coupled receptors. *Comment on Nat. Struct. Mol. Biol.*, 12:320–326, **2005**.

[53] W. Liu, M. Eilers, A. B. Patel, and S. O. Smith. Helix Packing Moments Reveal Diversity and Conservation in Membrane Protein Structure. *J. Mol. Biol.*, 337(3):713–729, **2004**.

[54] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano. Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science*, 289(5480):739–745, **2000**.

[55] G. F. J. Salgado, A. V. Struts, K. Tanaka, N. Fujioka, K. Nakanishi, and M. F. Brown. Deuterium NMR Structure of Retinal in the Ground State of Rhodopsin. *Biochemistry*, 43(40):12819–12828, **2004**.

[56] T. Okada, M. Sugihara, A. N. Bondar, M. Elstner, P. Entel, and V. Buss. The retinal conformation and its environment in rhodopsin in light of a new 2.2 AA;crystal structure. *J. Mol. Biol.*, 342(2):571–583, **2004**.

[57] T. P. Dryja, T. L. McGee, E. Reichel, L. B. Hahn, G. S. Cowley, D. W. Yandell, M. A. Sandberg, and E. L. Berson. A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. *Nature*, 343(6265):364–366, **1990**.

[58] A. M. Dizhoor, M. L. Woodruff, E. V. Olshevskaya, M. C. Cilluffo, M. C. Cornwall, P. A. Sieving, and G. L. Fain. Night Blindness and the Mechanism of Constitutive Signaling of Mutant G90D Rhodopsin. *J. Neurosci.*, 28(45):11662–11672, **2008**.

[59] R. B. Innis, F. M. Correa, and S. Synder. Carazolol, an extremely potent beta-adrenergic blocker: binding to beta-receptors in brain membranes. *Life Sci.*, 24(24):2255–2264, **1979**.

[60] M. R. Tayler. Pharmacogenetics of the human beta-adrenergic receptors. *Pharmacogenomics J.*, 7(1):29–37, **2007**.

[61] E. S. Burstein, T. A. Spalding, and M. R. Brann. The second intracellular loop of the m5 muscarinic receptor is the switch which enables G-protein coupling. *J. Biol. Chem.*, 273(38):24322–24327, **1998**.

[62] J. Sawynok and X. J. Liu. Adenosine in the spinal cord and periphery: release and regulation of pain. *Prog. Neurobiol.*, 69(5):313–340, **2003**.

[63] S. Lahiri, C. H. Mitchell, D. Reigada, A. Roy, and N. S. Cherniack. Purines, the carotid body and respiration. *Respir. Physiol. Neurobiol.*, 157(1):123–129, **2007**.

[64] R. Basheer, R. E. Strecker, M. M. Thakkar, and R. W. McCarley. Adenosine and sleep-wake regulation. *Prog. Neurobiol.*, 73(6):379–396, **2004**.

[65] B. Wu, E. Y. Chien, C. D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. C. Bi, D. J. Hamel, P. Kuhn, T. M. Handel, V. Cherezov, and R. C. Stevens. Structure of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science*, 330(6007):1066–1071, **2010**.

[66] M. Baggiolini. Chemokines and leuckocyte traffic. *Nature*, 392(6676):565–568, **1998**.

[67] B. Moser, M. Wolf, A. Walz, and P. Loetscher. Chemokines: multiple levels of leukocyte migration control. *Trends Immunol.*, 25(2):75–84, **2004**.

[68] C. R. Mackay. The chemokines: immunology's high impact factors. *Nat. Immunol.*, 2(2):95–101, **2001**.

[69] E. Y. Chien, W. Liu, Q. Zhao, V. Katritch, G. W. Han, M. A. Hanson, L. Shi, A. H. Newman, J. A. Javitch, V. Cherezov, and S. R. C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science*, 330(6007):1091–1095, **2010**.

[70] T. Shimamura, M. Shiroishi, S. Weyand, H. Tsujimoto, G. Winter, V. Katritch, R. Abagyan, V. Cherezov, W. Liu, G. W. Han, T. Kobayashi, R. C. Stevens, and S. Iwata. Structure of the human histamine H1 receptor complex with doxepin. *Nature*, 475(7354):65–70, **2011**.

[71] S. J. Hill, C. R. Ganellin, H. Timmerman, J. C. Schwartz, N. P. Shankley, J. M. Young, W. Schunack, R. Levi, and H. Haas. International Union of Pharmacology. XIII. Classification of Histamine Receptors. *Pharmacol. Rev.*, 49(3):253–278, **1997**.

[72] J. Kopp, L. Bordoli, J. N. D. Battey, F. Kiefer, and T. Schwede. Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, 69 Suppl 8:38–56, **2007**.

[73] A. Martinelli and T. Tuccinardi. An overview of recent developments in GPCR modelling: methods and validation. *Expert Opin. Drug Discovery*, 1(5):459–476, **2006**.

[74] E. Filizola and M. Filizola. Advances in the development and application of computational methodologies for structural modeling of G-protein-coupled receptors. *Expert Opin. Drug Discovery*, 3(3):343–355, **2008**.

[75] A. M. Lesk and C. H. Chothia. The Response of Protein Structures to Amino-Acid Sequence Changes. *Philos. Trans. R. Soc.*, 317(1540):345–356, **1986**.

[76] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, 29:291–325, **2000**.

[77] B. Kneissl, B. Leonhardt, A. Hildebrandt, and C. S. Tautermann. Revisiting automated G-protein coupled receptor modeling: the benefit of additional template structures for a neurokinin-1 receptor model. *J. Med. Chem.*, 52(10):3166–3173, **2009**.

[78] M. Nowak, M. Kolaczkowski, M. Pawlowski, and A. J. Bojarski. Homology modeling of the serotonin 5-HT1A receptor using automated docking of bioactive compounds with defined geometry. *J. Med. Chem.*, 49(1):205–214, **2006**.

[79] J. N. Pennefather, A. Lecci, M. L. Candenas, E. Patak, F. M. Pinto, and C. A. Maggi. Tachykinins and tachykinin receptors: a growing family. *Life Sci.*, 74(12):1445–1463, **2004**.

[80] R. Patacchini and C. A. Maggi. Peripheral tachykinin receptors as targets for new drugs. *Eur. J. Pharmacol.*, 429(1-3):13–21, **2001**.

[81] M. A. Cascieri, L. L. Shiao, S. G. Mills, M. Maccoss, C. J. Swain, H. Yu, E. Ber, S. Sadowski, M. T. Wu, C. D. Strader, and T. M. Fong. Characterization of the interaction of diacylpiperazine antagonists with the human neurokinin-1 receptor-identification of a common binding-site for structurally dissimilar antagonists. *Mol. Pharmacol.*, 27(4):660–665, **1995**.

[82] A. Evers and G. Klebe. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *J. Med. Chem.*, 47(22):5381–5392, **2004**.

[83] MOE - Molecular Operating Environment. Chemical Computing Group, Montreal, **2007**.

[84] GLIDE: Grid-based Ligand Docking with Energetics. Schroedinger: New York, **2004**.

[85] Symyx Draw 3.1. Symyx Technologies Inc.: Santa Clara, CA, **2008**.

[86] G. J. Barton. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.*, 6(1):37–40, **1993**.

[87] A. Moll, A. Hildebrandt, H. P. Lenhof, and O. Kohlbacher. Ballview: An object-oriented molecular visualization and modeling framework. *J. Comput. Aided Mol. Des.*, 19(11):791–800, **2005**.

[88] O. Kohlbacher and H. P. Lenhof. BALL - rapid software prototyping in computational molecular biology. *Bioinformatics*, 16(9):815–824, **2000**.

[89] R. Gentleman and R. Ihaka. R: A Language and Environment for Statistical Computing. Institute for Statistics and Mathematics, Vienna University of Economics and Business: Vienna, Austria, **2008**.

[90] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234(3):779–815, **1993**.

[91] R. M. Snider, J. W. Constantine, J. A. Lowe, K. P. Longo, W. S. Lebel, H. A. Woody, S. E. Drozda, M. C. Desai, F. J. Vinick, R. W. Spencer, and H. J. Hess. A potent nonpeptide antagonist of the substance-P (NK1) receptor. *Science*, 251(4992):435–437, **1991**.

[92] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–1749, **2004**.

[93] T. M. Fong, H. Yu, M. A. Cascieri, D. Underwood, C. J. Swain, and C. D. Strader. Interaction of Glutamine-165 in the 4th transmembrane segment of the human Neurokinin-1 receptor with quinuclidine antagonist. *J. Biol. Chem.*, 269(21):14957–14961, **1994**.

[94] T. M. Fong, M. A. Cascieri, H. Yu, A. Bansal, C. Swain, and C. D. Strader. Amino Aromatic Interaction between Histidine-197 of the Neurokinin-1 receptor and CP-96345. *Nature*, 362(6418):350–353, **1993**.

[95] T. M. Fong, H. Yu, M. A. Cascieri, D. Underwood, C. J. Swain, and C. D. Strader. The role of Histidine-265 in the antagonist binding to the Neurokinin-1 receptor. *J. Biol. Chem.*, 269(4):2728–2732, **1994**.

[96] T. M. Fong, H. Yu, R. R. C. Huang, M. A. Cascieri, and C. J. Swain. Relative contribution of polar interactions and conformational compatibility to the binding of neurokinin-1 receptor antagonists. *Mol. Pharm.*, 50(6):1605–1611, **1996**.

[97] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, **2001**.

[98] T. K. Attwood and J. B. C. Findlay. Fingerprinting G-Protein coupled rececptors. *Protein Eng.*, 7(2):195–203, **1994**.

[99] F. Menzaghi, D. P. Behan, and D. T. Chalmers. Constitutively activated G protein-coupled receptors: a novel approach to CNS drug discovery. *Curr. Drug Targets CNS Neurol. Disord.*, 42(1):105–121, **2002**.

[100] F. Fanelli and P. G. De Benedetti. Computational Modeling approaches to structure-function analysis of G protein-coupled receptors. *Chem. Rev.*, 105(9):3297–3351, **2005**.

[101] T. Mirzadegan, G. Benko, S. Filipek, and K. Palczewski. Sequence analyses of G-protein-coupled receptors: Similarities to rhodopsin. *Biochemistry*, 42(10):2759–2767, **2003**.

[102] J. M. Baldwin, G. F. X. Schertler, and V. M. Unger. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.*, 272(1):144–164, **1997**.

[103] K. Hofmann and W. Stoffel. TMbase - A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler*, 374(166), **1993**.

[104] J. S. Surgand, J. Rodrigo, E. Kellenberger, and D. Rognan. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins: Struct., Func., Bioinf.*, 62(2):509–538, **2006**.

[105] C. Bissantz, P. Bernard, M. Hibert, and D. Rognan. Protein-based virtual screening of chemical databases. II. Are homology models of G-protein coupled receptors suitable targets? *Proteins*, 50(1):5–25, **2003**.

[106] M. Nowak, M. Kolaczkowski, M. Pawlowski, and A. J. Bojarski. Homology modeling of the serotonin 5-HT1A receptor using automated docking of bioactive compounds with defined geometry. *J. Med. Chem.*, 49(1):205–214, **2006**.

[107] P. Ferrara and E. Jacoby. Evaluation of the utility of homology models in high throughput docking. *J. Mol. Model.*, 13(8):897–905, **2007**.

[108] S. Bhattacharya, S. E. Hall, H. Li, and N. Vaidehi. Ligand-stabilized conformational states of human beta(2) adrenergic receptor: Insight into G-protein-coupled receptor activation. *Biophys. J.*, 94(6):2027–2042, **2008**.

[109] C. Bissantz, C. Schalon, W. Guba, and M. Stahl. Focused library design in GPCR projects on the example of 5-HT2c agonists: Comparison of structure-based virtual screening with ligand-based search methods. *Proteins*, 61(4):938–952, **2005**.

[110] B. Rost, P. Fariselli, and R. Casadio. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci*, 5(8):1704–1718, **1996**.

[111] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, 305(3):567–580, **2001**.

[112] M. J. Hwang, T. P. Stockfisch, and A. T. Hagler. Derivation and characterization of a class II force field, CFF93, for the alkyl functional group and alkane molecules. *J. Am. Chem. Soc.*, 116:2515–2525, **1994**.

[113] D. Ernst. Identification and Similarity Measurement of GPCR Binding Pockets. Bachelor's Thesis, Saarland University, **2011**.

[114] C. Cole, J. D. Barber, , and G. J. Barton. The Jpred 3 secondary structure prediction server. *Nucl. Acids Res.*, 36(Web Server issue):W197–W201, **2008**.

[115] W. Zhou, C. Flanagan, J. A. Ballesteros, K. Konvicka, J. S. Davidson, H. Weinstein, R. P. Millar, and S. C. Sealfon. A reciprocal mutation supports helix 2 and helix 7 proximity in the gonadotropinreleasing hormone receptor. *Mol. Pharmacol.*, 45(2):165–170, **1994**.

[116] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105–132, **1982**.

[117] D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, 179(1):125–142, **1984**.

[118] J. Janin. Surface and inside volumes in globular proteins. *Nature*, 277(5696):491–492, **1979**.

[119] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus. A simple method for displaying the hydropathic character of a protein. *Science*, 229(4716):834–838, **1985**.

[120] J. L. Cornette, K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, 195(3):659–685, **1987**.

[121] D. M. Engelman, T. A. Steitz, , and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.*, 15:321–353, **1986**.

[122] T. P. Hopp and K. R. Woods. A computer program for predicting protein antigenic determinants. *Mol. Immunol.*, 20(4):483–489, **1983**.

[123] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256(3):623–644, **1996**.

[124] T. Thies. Optimierung von Helix-Helix Interaktionen bei G-Protein-gekoppelten Rezeptoren. Master's Thesis, Wilhelm-Schickard Institut für Informatik, **2008**.

[125] M. Lundy and A. Mees. Convergence of an Annealing Algorithm. *Math. Prog.*, 34:111–124, **1986**.

[126] D. C. Liu and J. Nocedal. On the Limited Memory Method for Large Scale Optimization. *Math. Program. B.*, 45(3):503–528, **1989**.

[127] A. Rurainski. *Optimization in bioinformatics*. PhD thesis, Saarland University, **2010**.

[128] M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics*. Springer, **2007**.

[129] A. R. van Buuren and H. J. Berendsen. Molecular dynamics simulation of the stability of a 22-residue alpha-helix in water and 30% trifluoroethanol. *Biopolymers*, 33(8):1159–1166, **1993**.

[130] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, 4(3):435–447, **2008**.

[131] A. Lund. Analyzing Molecular Dynamic Trajectories of Peptides by means of GPCR Helices. Bachelor's Thesis, Johannes Gutenberg-University Mainz, **2011**.

[132] T. M. Yi and E. S. Lander. Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, 232(4):1117–1129, **1993**.

[133] A. A. Salamov and V. V. Solovyev. Protein secondary structure prediction using local alignments. *J. Mol. Biol.*, 268(1):31–36, **1997**.

[134] N. Qian and T. J. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, 202(4):865–884, **1998**.

[135] G. E. Tusnády and I. Simon. The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17(9):849–850, **2001**.

[136] H. Kim and H. Park. Protein secondary structure prediction based on an improved support vector machines approach. *Proteins*, 16(8):553–560, **2003**.

[137] R. Bondugula and D. Xu. MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins*, 66(3):664–670, **2007**.

[138] W. Pirovano and J. Heringa. Protein secondary structure prediction. *Methods Mol. Biol.*, 609:327–348, **2010**.

[139] S. Yohannan, S. Faham, D. Yang, J. P. Whitelegge, and J. U. Bowie. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.*, 101(4):959–963, **2004**.

[140] I. Rigoutsos, P. Riek, R. M. Graham, and J. Novotny. Structural details (kinks and non-alpha conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern descriptors. *Nucleic Acids Res.*, 31(15):4625–4631, **2003**.

[141] S. E. Hall, K. Roberts, and N. Vaidehi. Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction. *J. Mol. Graph. Model.*, 27(8):944–950, **2009**.

[142] D. N. Langelaan, M. Wieczorek, C. Blouin, and J. K. Rainey. Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. *J. Chem. Inf. Model.*, 50(12):2213–2220, **2010**.

[143] A. D. Meruelo, I. Samish, and J. U. Bowie. TMKink: A method to predict transmembrane helix kinks. *Protein Sci.*, 20(7):1256–1264, **2011**.

[144] B. Kneissl, S. C. Müller, C. S. Tautermann, and A. Hildebrandt. String Kernels and High-Quality Data Set for Improved Prediction of Kinked Helices in α-Helical Membrane Proteins. *J. Chem. Inf. Mod.*, 51(11):3017–3025, **2011**.

[145] G. Wang and R. L. Dunbrack. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.*, 33(Web server issue):W94–W98, **2005**.

[146] G. E. Tusnády, Z. Dosztányi, and I. Simon. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, 21(7):1276–1277, **2005**.

[147] R. P. Riek, I. Rigoutsos, J. Novotny, and R. M. Graham. Non-alpha-helical elements modulate polytopic membrane protein architecture. *J. Mol. Biol.*, 306(2):349–362, **2001**.

[148] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2(6):559–572, **1901**.

[149] M. Bansal, S. Kumar, and R. Velavan. HELANAL: a program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.*, 17(5):811–819, **2000**.

[150] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, 32:2319–2327, **2011**.

[151] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20:273–297, **1995**.

[152] J. Gubbi, A. Shilton, M. Parker, and M. Palaniswami. Protein topology classification using two-stage support vector machines. *Genome Inform.*, 17(2):259–269, **2006**.

[153] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.*, 7:564–575, **2002**.

[154] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, **2004**.

[155] Y. Zhao, C. Pinilla, D. Valmori, R. Martin, and R. Simon. Application of support vector machines for T-cell epitopes prediction. *Bioinformatics*, 19(15):1978–1984, **2003**.

[156] J. Shi, F.; Huang. Prediction of T-cell Epitopes Using Support Vector Machine and Similarity Kernel. *CIS*, 1:604–608, **2005**.

[157] H. Saigo, J. P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, **2004**.

[158] C.-C. Chang and C.-J. Lin. LIBSVM: a Library for Support Vector Machines. *ACM Trans. Sys. Tech.*, 27:1–27, **2011**.

[159] A. Hildebrandt, A. K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. C. Toussaint, A. Moll, D. Stöckel, S. Nickels, S. C. Mueller, H. P. Lenhof, and O. Kohlbacher. BALL - biochemical algorithms library 1.3. *BMC Bioinformatics*, 11:531, **2010**.

[160] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer: New York, **2003**.

[161] R. Kohavi and F. Provost. *Glossary of Terms. In Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, volume 30, pages 271–274. Kluwer Academic Publishers: New York, **1998**.

[162] G. von Heijne. Proline kinks in transmembrane alpha-helices. *J. Mol. Biol.*, 218(3):499–503, **1991**.

[163] V. Geetha. Distortions in protein helices. *Int. J. Biol. Macromol.*, 19(2):81–89, **1996**.

[164] S. Pal, A. Heifetz, R. J. Law, A. Kahrs, T. Hesterkamp, J. Madden, A. Davenport, A. Parkes, M. Mazanetz, D. Hallet, and M. Whittaker. Hierarchical GPCR modelling; application to Bradykinin 1 Receptor (B1R), Histamine 3 (H3) receptor and Melanin-Concentrating Hormone receptor 1 (MCHR1). In *High Resolution Neuropharmacology meeting*, **2010**.

[165] A. Baldauf. The Applicability of Artificial Evolution for modeling transmembrane helices by means of GPCRs. Master's Thesis, Saarland University, **2011**.

[166] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77(4):778–95, **2009**.

[167] M. E. Rose. *Elementary Theory of Angular Momentum*. Wiley, New York, **1957**.

[168] E. B. Dam, M. Koch, and M. Lillholm. Quaternions, Interpolation and Animation. Technical Report MSU-CSE-00-2, Department of Computer Science, University of Copenhagen, **1998**.

[169] N. Vaidehi, W. B. Floriano, R. Trabanino, S. E. Hall, P. Freddolino, and E. J. Choi. Prediction of structure and function of G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA*, 99(20):12622–12627, **2002**.

[170] C. L. Worth, G. Kleinau, and G. Krause. Comparative sequence and structural analyses of G-protein-coupled receptor crystal structures and implications for molecular models. *PloS one*, 4(9):e7011, **2009**.

[171] V. Katritch, V. Cherezov, and R. C. Stevens. Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol. Sci.*, 33(1):17–27, **2012**.